

# SPLITTING FORWARD-BACKWARD PENALTY SCHEME FOR CONSTRAINED VARIATIONAL PROBLEMS

MARC-OLIVIER CZARNECKI, NAHLA NOUN, AND JUAN PEYPOUQUET

**ABSTRACT.** We study a forward backward splitting algorithm that solves the variational inequality

$$Ax + \nabla\Phi(x) + N_C(x) \ni 0$$

where  $\mathcal{H}$  is a real Hilbert space,  $A : \mathcal{H} \rightrightarrows \mathcal{H}$  is a maximal monotone operator,  $\Phi : \mathcal{H} \rightarrow \mathbf{R}$  is a smooth convex function, and  $N_C$  is the outward normal cone to a closed convex set  $C \subset \mathcal{H}$ . The constraint set  $C$  is represented as the intersection of the sets of minima of two convex penalization function  $\Psi_1 : \mathcal{H} \rightarrow \mathbf{R}$  and  $\Psi_2 : \mathcal{H} \rightarrow \mathbf{R} \cup \{+\infty\}$ . The function  $\Psi_1$  is smooth, the function  $\Psi_2$  is proper and lower semicontinuous. Given a sequence  $(\beta_n)$  of penalization parameters which tends to infinity, and a sequence of positive time steps  $(\lambda_n)$ , the algorithm

$$(SFBP) \quad \begin{cases} x_1 & \in \mathcal{H}, \\ x_{n+1} & = (I + \lambda_n A + \lambda_n \beta_n \partial\Psi_2)^{-1}(x_n - \lambda_n \nabla\Phi(x_n) - \lambda_n \beta_n \nabla\Psi_1(x_n)), \quad n \geq 1. \end{cases}$$

performs forward steps on the smooth parts and backward steps on the other parts. Under suitable assumptions, we obtain weak ergodic convergence of the sequence  $(x_n)$  to a solution of the variational inequality. Convergence is strong when either  $A$  is strongly monotone or  $\Phi$  is strongly convex. We also obtain weak convergence of the whole sequence  $(x_n)$  when  $A$  is the subdifferential of a proper lower-semicontinuous convex function. This provides a unified setting for several classical and more recent results, in the line of historical research on continuous and discrete gradient-like systems.

**Key words:** constrained convex optimization; forward-backward algorithms; hierarchical optimization; maximal monotone operators; penalization methods; variational inequalities.

**AMS subject classification.** 37N40, 46N10, 49M30, 65K05, 65K10, 90B50, 90C25.

## 1. INTRODUCTION

In 1974, R. Bruck [15] showed that the trajectories of the steepest descent system

$$\dot{x}(t) + \partial\Phi(x(t)) \ni 0$$

minimize the convex, proper, lower-semicontinuous potential  $\Phi$  defined on a real Hilbert space  $\mathcal{H}$ . They weakly converge to a point in the minima of  $\Phi$  and the potential decreases along the trajectory toward its minimal value, provided  $\Phi$  attains its minimum. When the semigroup is generated by the differential inclusion

$$\dot{x}(t) + A(x(t)) \ni 0$$

with a maximal monotone operator  $A$  from  $\mathcal{H}$  to  $\mathcal{H}$ , J.-B. Baillon and H. Brézis [8] provided in 1976 the convergence in average to an equilibrium of  $A$ . These results are sharp. If the operator is a rotation in  $\mathbf{R}^2$ , the trajectories do not converge, except the stationary one. J.-B. Baillon [6], provided an example in 1978, where the trajectories of the steepest descent system do not strongly converge, although they weakly converge. In some sense, his example is an extension to a Hilbert space of the rotation in  $\mathbf{R}^2$ . The keytool for the proof of these two results is Opial's lemma [30] that gives weak convergence without a priori knowledge of the limit. In 1996, H. Attouch and R.

Cominetti [2] coupled approximation methods with the steepest descent system, in particular by adding a Tikhonov regularizing term:

$$\dot{x}(t) + \partial\Psi(x(t)) + \varepsilon(t)x(t) \ni 0.$$

The parameter  $\varepsilon$  tends to zero and the potential  $\Psi$  satisfies usual assumptions. As it yields the steepest descent system for  $\varepsilon = 0$ , one can expect the trajectories to weakly converge. The striking point of their results is the strong convergence of the trajectories, when  $\varepsilon$  tends to 0 slowly enough, that is  $\varepsilon$  does not belong to  $l^1$ . Then the strong limit is the point of minimal norm among the minima of  $\Psi$ . This seems rather surprising at first: without a regularizing term  $\varepsilon(t)x(t)$ , we know that we only have weak convergence, and with the regularizing term, convergence is strong. We propose the following explanation: set  $\Phi(x) = \frac{1}{2}\|x\|^2$  so that the regularizing term writes  $\varepsilon(t)x(t) = \varepsilon(t)\nabla\Phi(x(t))$ . Then, by a change of time, valid for  $\varepsilon \notin l^1$ , see [3], we can reformulate the system as a penalized system

$$\dot{x}(t) + \nabla\Phi(x(t)) + \beta(t)\partial\Psi(x(t)) \ni 0,$$

with a parameter  $\beta$  that tends to infinity. But now we are looking at a steepest descent system for the strongly convex function  $\Phi$  with a penalization potential  $\Psi$ , possibly equal to 0. And it is known that the trajectories of the steepest descent system strongly converge when the potential is strongly convex. In 2004, A. Cabot [18] generalized part of Attouch and Cominetti's result to the case of a strongly convex potential. The penalization acts as a constraint and forces the limit to belong to the minima of  $\Psi$ . It appeared natural to add a penalization, rather than a perturbation or a regularization, to the first order differential inclusion with a maximal monotone operator, and to the steepest descent system with a -not necessarily strongly- convex potential. Moreover, penalization methods enjoy great practical interest when others, such as projection methods, encounter intrinsic implementation difficulties, for example when the constraint set is given by nonlinear inequalities. H. Attouch and M.-O. Czarnecki [3] showed in 2010 that the trajectories of

$$\dot{x}(t) + A(x(t)) + \beta(t)\partial\Psi(x(t)) \ni 0.$$

weakly converge in average to a constrained equilibrium

$$x_\infty \in (A + N_C)^{-1}(0),$$

that convergence is strong when  $A$  is strongly monotone, and in the subdifferential case, the trajectories of

$$\dot{x}(t) + \partial\Phi(x(t)) + \beta(t)\partial\Psi(x(t)) \ni 0.$$

weakly converge to a constrained minimum

$$x_\infty \in \operatorname{argmin}\{\Phi | \operatorname{argmin}\Psi\}.$$

Besides assuming the parameter  $\beta$  to tend to  $+\infty$ , their main assumption relates the geometry of the penalization potential  $\Psi$  to the growth rate of the penalization parameter  $\beta$ , namely

$$\int_0^{+\infty} \beta(t) \left[ \Psi^* \left( \frac{p}{\beta(t)} \right) - \sigma_C \left( \frac{p}{\beta(t)} \right) \right] dt < +\infty$$

for every  $p$  in the range of  $N_C$ , the normal cone to  $C = \operatorname{argmin}\Psi$ . Here  $\Psi^*$  denotes the Fenchel conjugate of  $\Psi$  and  $\sigma_C$  the support function of  $C$ . A detailed analysis of the condition is done in [3]. Let us just mention that, when  $\Psi = \frac{1}{2}\operatorname{dist}_C^2$ , it reduces to  $\int_0^{+\infty} \frac{1}{\beta(t)} dt < +\infty$ . When  $\Psi = 0$ , then  $C = \mathcal{H}$  and the only  $p$  in the normal cone is 0, and the condition is fulfilled. So one recovers the results of Bruck and of Baillon and Brézis.

**Discretization.** In order to compute the trajectories of the system, and obtain workable algorithms, we need to discretize it. The implicit discretization of unpenalized first order system is the famous proximal algorithm, well studied from the mid seventies (B. Martinet [25] and [26], R.T. Rockafellar [35], H. Brézis and P.L. Lions. [14],...):

$$x_{n+1} = (I + \lambda_n A)^{-1} x_n.$$

Following the same path, in 2011, Attouch, Czarnecki and Peypouquet [4] discretized the penalized continuous system implicitly to obtain the backward algorithm:

$$x_{n+1} = (I + \lambda_n A + \lambda_n \beta_n \partial \Psi)^{-1} x_n.$$

They provide weak convergence in average to a constrained equilibrium  $x_\infty \in (A + N_C)^{-1}(0)$ , strong convergence when  $A$  is strongly monotone, and weak convergence in the subdifferential case. A basic assumption on the time step is  $(\lambda_n) \notin l^1$ , which corresponds to  $t \rightarrow +\infty$  in continuous time. The key assumption is the discrete counterpart of the assumption in continuous time:

$$\sum_{n=1}^{\infty} \lambda_n \beta_n \left[ \Psi^* \left( \frac{z}{\beta_n} \right) - \sigma_C \left( \frac{z}{\beta_n} \right) \right] < \infty$$

for every  $p$  in the range of  $N_C$ . Again, in the case where  $\Psi = 0$ , one recovers classical results for the proximal algorithm. The main drawback of the backward algorithm is the cost of every step in the computation of the discrete trajectory. The explicit discretization of the steepest descent system can be traced back to A. Cauchy [20] in 1847, who gave indeed the idea of the discrete gradient method

$$x_{n+1} = x_n - \lambda_n \nabla \Phi(x_n),$$

with no proof of convergence, and before the continuous steepest descent system. J. Peypouquet [33] discretized the continuous penalized system explicitly in 2012 to obtain a forward algorithm, in a regular setting

$$x_{n+1} = x_n - \lambda_n \nabla \Phi(x_n) - \lambda_n \beta_n \nabla \Psi(x_n).$$

He shows weak convergence of the trajectories to a constrained minimum

$$x_\infty \in \operatorname{argmin}\{\Phi | \operatorname{argmin}\Psi\}$$

provided the gradient of the potential  $\Phi$  is Lipschitz continuous. Together with the key assumption described before relating the Fenchel conjugate of  $\Psi$  and the sequences  $(\lambda_n)$  and  $(\beta_n)$ , he requires an assumption combining a bound on these sequences and the Lipschitz constant of  $\nabla \Psi$ . It slightly differs, and is a consequence of

$$\limsup_{n \rightarrow \infty} \lambda_n \beta_n < \frac{2}{L_{\nabla \Psi}}.$$

To make things short, forward algorithms are more performing, but require more regularity assumptions and convergence is more complicated to prove, while backward algorithms apply to more general cases, convergence is easier to prove but they are not so efficient. As the constraint set can possibly be described by a regular penalization function, the next idea, developed in [5], is to perform a forward step on the regular part  $\Psi$ , and a backward step on the other part to obtain the forward-backward algorithm:

$$x_{n+1} = (I + \lambda_n A)^{-1} (x_n - \lambda_n \beta_n \nabla \Psi(x_n)).$$

We obtain again weak convergence in average to a constrained equilibrium  $x_\infty \in (A + N_C)^{-1}(0)$ , strong convergence when  $A$  is strongly monotone, and weak convergence in the subdifferential case. Together with the key summability assumption relating  $\Psi$  and the parameters  $(\lambda_n)$  and  $(\beta_n)$ , we

assume regularity of the function  $\Psi$ , that it is differentiable with a Lipschitz gradient. We need the same assumption on  $(\lambda_n)$ ,  $(\beta_n)$ , and the Lipschitz constant of  $\nabla\Psi$ :

$$\limsup_{n \rightarrow \infty} \lambda_n \beta_n < \frac{2}{L_{\nabla\Psi}}.$$

The bound is strict in general, and is close to being sharp: equality, with a precise potential  $\Psi$ , corresponds to a result of Passty [31] on alternate algorithms. A detailed analysis is given in [5].

**Regularity based splitting.** We now have three different algorithms depending on the regularity of the data: backward algorithms, forward algorithms, forward-backward algorithms. Convergence holds under the same summability assumption involving the Fenchel conjugate of  $\Psi$ , and similar regularity assumptions to perform a forward step.

What if the maximal monotone operator has a regular part, and if the penalization potential decomposes with a regular part? Can we guarantee convergence if we perform forward steps on the regular parts, while keeping the backward step on the other parts? Can we provide a unified setting for the previous algorithms?

Let  $A$  be a maximal monotone operator, let  $\Phi$ ,  $\Psi_1$ ,  $\Psi_2$  be convex proper lower semicontinuous potentials. The functions  $\Phi$  and  $\Psi_2$  are defined everywhere and differentiable with a Lipschitz gradient. Set  $C = \text{Argmin}\Psi_1 \cap \text{Argmin}\Psi_2$ , which corresponds to the decomposition of the penalization function as the sum of a smooth part and a general part, and assume that  $C$  is not empty. The penalized system

$$(1) \quad \dot{x}(t) + (A + \nabla\Phi)(x(t)) + \beta(t)(\partial\Psi_1 + \nabla\Psi_2)(x(t)) \ni 0.$$

allows to solve the variational inequality

$$(2) \quad 0 \in Ax + \nabla\Phi(x) + N_C(x).$$

We discretize this last continuous penalized system by making a forward step on the regular parts  $\Phi$  and  $\Psi_2$ , and a backward step on  $A$  and  $\Psi_1$ . Given a positive sequence  $(\beta_n)$  of *penalization parameters*, along with a positive sequence  $(\lambda_n)$  of *step sizes*, we consider the *splitting forward-backward penalty algorithm* (SFBP), defined as follows:

$$(SFBP) \quad \begin{cases} x_1 & \in \mathcal{H}, \\ x_{n+1} & = (I + \lambda_n A + \lambda_n \beta_n \partial\Psi_2)^{-1}(x_n - \lambda_n \nabla\Phi(x_n) - \lambda_n \beta_n \nabla\Psi_1(x_n)), \quad n \geq 1. \end{cases}$$

Under suitable assumptions, including the expected geometrical condition involving the Fenchel conjugate of the penalization potential and the expected relationship between the parameters and the Lipschitz constant of  $\nabla\Psi_1$ , we prove that, as  $n \rightarrow \infty$ , the sequences generated by the (SFBP) algorithm converge to a constrained equilibrium in  $S = (A + \nabla\Phi + N_C)^{-1}0$

- i) weakly in average if  $A$  is any maximal monotone operator (Theorem 1);
- ii) strongly if  $A$  is strongly monotone, or if  $\Phi$  is strongly convex (Theorem 2);
- iii) weakly if  $A$  is the subdifferential of a proper, lower-semicontinuous and convex function (Theorem 3).

Besides its applicability to problems that combine smooth and nonsmooth features, the (SFBP) algorithm allows us to study, in a unified framework, the classical and more recent methods to solve constrained variational inequalities.

If  $\Phi = \Psi_1 = \Psi_2 \equiv 0$ , we recover the *proximal point algorithm*. If  $\Phi = \Psi_1 \equiv 0$ , the (SFBP) algorithm corresponds to the purely implicit *prox-penalization algorithm* from [4]. The *gradient method*, is recovered in the case where  $A \equiv 0$  and  $\Psi_1 = \Psi_2 \equiv 0$ . If  $A \equiv 0$  and  $\Psi_2 \equiv 0$ , we obtain the purely explicit *diagonal gradient scheme* from [33]. We get the *forward-backward splitting* (a combination of the *proximal point algorithm* and of the *gradient method*, see [31]) if  $\Psi_1 = \Psi_2 \equiv 0$ . The case  $\Phi = \Psi_2 \equiv 0$  gives a semi-implicit penalty splitting method studied in [5] and [29].

**Organization of the paper.** The paper is organized as follows: We begin by describing and commenting the hypothesis, and stating the main theoretical results, in Section 2. Next, in Section 3, we present several special cases, and compare them with classical and more recent methods to solve constrained optimization problems. In particular, this work extends and unifies some previous developments progressively achieved by our work group. In Section 4, we describe a model for the sparse-optimal control of a linear system of ODE's. It illustrates how the decomposition of objective potential and penalization naturally arises. Finally, we present the proofs in several steps in Section 5.

## 2. MAIN RESULTS

Let  $\mathcal{H}$  be a real Hilbert space. We first recall some facts about convex analysis and maximal monotone operator theory. Let  $\Gamma_0(\mathcal{H})$  be the set of all proper (not identically equal to  $+\infty$ ) lower-semicontinuous convex functions from  $\mathcal{H}$  to  $\mathbf{R} \cup \{+\infty\}$ . Given  $F \in \Gamma_0(\mathcal{H})$  and  $x \in \mathcal{H}$ , the *subdifferential* of  $F$  at  $x$  is the set

$$\partial F(x) = \{x^* \in \mathcal{H} : F(y) \geq F(x) + \langle x^*, y - x \rangle \text{ for all } y \in \mathcal{H}\}.$$

Given a nonempty closed convex set  $C \subset \mathcal{H}$ , its *indicator function* is defined as  $\delta_C(x) = 0$  if  $x \in C$  and  $+\infty$  otherwise. The *normal cone* to  $C$  at  $x$  is

$$N_C(x) = \{x^* \in \mathcal{H} : \langle x^*, c - x \rangle \leq 0 \text{ for all } c \in C\}$$

if  $x \in C$  and  $\emptyset$  otherwise. Observe that  $\partial \delta_C = N_C$ . A *monotone operator* is a set-valued mapping  $A : \mathcal{H} \rightarrow \mathcal{H}$  such that  $\langle x^* - y^*, x - y \rangle \geq 0$  whenever  $x^* \in Ax$  and  $y^* \in Ay$ . It is *maximal monotone* if its graph is not properly contained in the graph of any other monotone operator. It is convenient to identify a maximal monotone operator  $A$  with its graph, thus we equivalently write  $x^* \in Ax$  or  $[x, x^*] \in A$ . The inverse  $A^{-1} : \mathcal{H} \rightarrow \mathcal{H}$  of  $A$  is defined by  $x \in A^{-1}x^* \Leftrightarrow x^* \in Ax$ . It is still a maximal monotone operator. For any maximal monotone operator  $A : \mathcal{H} \rightarrow \mathcal{H}$  and for any  $\lambda > 0$ , the operator  $I + \lambda A$  is surjective by Minty's Theorem (see [13] or [34]). The operator  $(I + \lambda A)^{-1}$  is nonexpansive and everywhere defined. It is called the *resolvent* of  $A$  of index  $\lambda$ .

Finally recall that the subdifferential of a function in  $\Gamma_0(\mathcal{H})$  is maximal monotone.

**2.1. Assumptions.** Let  $A$  be a maximal monotone operator, let  $\Phi, \Psi_1, \Psi_2$  be convex proper lower semicontinuous potentials with

$$C = \text{Argmin} \Psi_1 \cap \text{Argmin} \Psi_2 \neq \emptyset.$$

The functions  $\Phi$  and  $\Psi_2$  are defined everywhere and differentiable with a Lipschitz gradient. The *Fenchel conjugate* of a proper, lower-semicontinuous and convex function  $F : \mathcal{H} \rightarrow \mathbf{R} \cup \{+\infty\}$  is the function  $F^* : \mathcal{H} \rightarrow \mathbf{R} \cup \{+\infty\}$  defined by

$$F^*(x^*) = \sup_{y \in \mathcal{H}} \{\langle y, x^* \rangle - F(y)\}$$

for each  $x^* \in \mathcal{H}$ . It is also proper, lower-semicontinuous and convex. Given a nonempty, closed and convex set  $C$ , its *support function*  $\sigma_C$  is defined as  $\sigma_C(x^*) = \sup_{c \in C} \langle x^*, c \rangle$  for  $x^* \in \mathcal{H}$ . Observe that  $\delta_C^* = \sigma_C$ . Notice also that  $x^* \in N_C(x)$  if, and only if,  $\sigma_C(x^*) = \langle x^*, x \rangle$ .

The main set of hypotheses is the following:

$$(H_0) \quad \begin{cases} i) & \mathbf{T} = A + \nabla \Phi + N_C \text{ is maximal monotone and } S = \mathbf{T}^{-1}(0) \neq \emptyset; \\ ii) & \nabla \Phi \text{ is } L_\Phi\text{-Lipschitz-continuous and } \nabla \Psi_1 \text{ is } L_{\Psi_1}\text{-Lipschitz-continuous}; \\ iii) & \text{For each } z \in N_C(\mathcal{H}), \quad \sum_{n=1}^{\infty} \lambda_n \beta_n \left[ (\Psi_1 + \Psi_2)^* \left( \frac{z}{\beta_n} \right) - \sigma_C \left( \frac{z}{\beta_n} \right) \right] < \infty; \\ iv) & \sum_{n=1}^{\infty} \lambda_n = +\infty, \quad \sum_{n=1}^{\infty} L_\Phi \lambda_n^2 < +\infty, \quad \text{and} \quad \limsup_{n \rightarrow \infty} (L_{\Psi_1} \lambda_n \beta_n) < 2. \end{cases}$$

We adress some remarks on Hypothesis  $(H_0)$  in order.

*On Part i).* It is a well-posedness and qualification condition ensuring that

$$(3) \quad \bar{u} \in S \quad \text{if, and only if,} \quad \langle w, u - \bar{u} \rangle \geq 0 \text{ for all } [u, w] \in \mathbf{T}.$$

If  $A$  is the subdifferential of a proper, lower-semicontinuous and convex function  $\Phi_2$ , the maximal monotonicity of  $\mathbf{T}$  implies

$$(4) \quad S = \text{Argmin}\{\Phi(x) + \Phi_2(x) : x \in C\}.$$

In this situation,  $S$  can be interpreted as the set of solutions of a hierarchical optimization problem, where  $\Phi + \Phi_2$  and  $\Psi_1 + \Psi_2$  are primary and secondary criteria, respectively. In this case, maximality of  $\mathbf{T}$  holds under some qualification condition, such as Moreau-Rockafellar or Attouch-Brézis.

*On Part ii).* It is standard for the convergence of gradient-related methods (see [11]).

*On Part iii).* It was introduced in [4], following [3]. The potentials  $\Psi_1$  and  $\Psi_2$  enter the algorithm only via their subdifferentials. Thus it is not a restriction to assume  $\min \Psi_1 = \min \Psi_2 = 0$ . Otherwise, one should replace  $\Psi_i$  by  $\Psi_i - \min \Psi_i$  in the corresponding statements. In the unconstrained case ( $\Psi_1 = \Psi_2 \equiv 0$ ), the condition is trivially satisfied since  $N_C(\mathcal{H}) = \{0\}$ ,  $\Psi^*(0) = 0$  and  $\sigma_{\mathcal{H}}(0) = 0$ . We refer to [4] for discussion and sufficient conditions. Note that the constraint set  $C$  is the set of minima of the potential  $\Psi_1 + \Psi_2$ , which leads naturally to an assumption on the Fenchel conjugate of the sum  $\Psi_1 + \Psi_2$ , and involving points in the normal cone  $N_C$ . In our setting, considering alternatively the two separate corresponding conditions on  $\Psi_1^*$  and  $\Psi_2^*$  would require extra qualification conditions.

*On Part iv).* The nonsummability condition in Part iv) is standard for the proximal point algorithm (see [14]) and gradient-related methods (see [11]). The second condition holds if either  $\Phi$  is affine (that is  $L_\Phi = 0$ ) or  $(\lambda_n)$  is in  $\ell^2$ . We write  $\limsup_{n \rightarrow \infty} (L_{\Psi_1} \lambda_n \beta_n) < 2$  rather than  $\limsup_{n \rightarrow \infty} \lambda_n \beta_n < \frac{2}{L_{\Psi_1}}$  to include the case where  $L_{\Psi_1} = 0$  ( $\Psi_1 \equiv 0$ ).

**2.2. Convergence results.** Take a sequence  $(x_n)$  generated by the *splitting forward-backward penalty algorithm* (SFBP):

$$(SFBP) \quad \begin{cases} x_1 & \in \mathcal{H}, \\ x_{n+1} & = (I + \lambda_n A + \lambda_n \beta_n \partial \Psi_2)^{-1}(x_n - \lambda_n \nabla \Phi(x_n) - \lambda_n \beta_n \nabla \Psi_1(x_n)), \quad n \geq 1, \end{cases}$$

which corresponds to the implicit-explicit discretization

$$\frac{x_n - x_{n+1}}{\lambda_n} - \nabla \Phi(x_n) - \beta_n \nabla \Psi_1(x_n) \in A x_{n+1} + \beta_n \partial \Psi_2(x_{n+1}).$$

of the penalized differential inclusion (1). We do not discuss in detail the existence of trajectories. Maximality of  $A + \beta_n \partial \Psi_2$  for all  $n \in \mathbf{N}$  is sufficient in view of Minty's theorem. Moreover, according the discussion in Subsection 2.3, it is possible to consider the above inclusion replacing the subdifferential operator by some enlargement, such as the  $\varepsilon$ -approximate subdifferential.

The kind of convergence depends on the nature of the operator  $A$ .

When  $A$  is any maximal monotone operator we prove the weak ergodic convergence of the algorithm to a point in  $S$ . More precisely, let  $(x_n)$  be a sequence generated by (SFBP) and let



$\tau_n = \sum_{k=1}^n \lambda_k$ . We define the following sequences of weighted averages:

$$(5) \quad z_n = \frac{1}{\tau_n} \sum_{k=1}^n \lambda_k x_k \quad \hat{z}_n = \frac{1}{\tau_n} \sum_{k=1}^n \lambda_k x_{k+1}.$$

Although quite similar, they converge under slightly different assumptions, as we shall see.

**Theorem 1.** *Assume that  $(H_0)$  holds. Then the sequence  $(\hat{z}_n)$  converges weakly to a point in  $S$  as  $n \rightarrow \infty$ . If we moreover assume that  $(\lambda_n) \in \ell^2$  then the sequence  $(z_n)$  converges weakly to a point in  $S$  as  $n \rightarrow \infty$ .*

Under further assumptions, it is possible to obtain strong or weak convergence of the whole sequence  $(x_n)$ . Recall that  $A$  is *strongly monotone* with parameter  $\alpha > 0$  if

$$\langle x^* - y^*, x - y \rangle \geq \alpha \|x - y\|^2$$

whenever  $x^* \in Ax$  and  $y^* \in Ay$ . The function  $\Phi$  is *strongly convex* if  $\nabla \Phi$  is strongly monotone. The set of zeros of a maximal monotone operator which is strongly monotone must contain exactly one element. We have the following:

**Theorem 2.** *Let  $(H_0)$  hold. If the operator  $A$  is strongly monotone, or if the potential  $\Phi$  is strongly convex, then every sequence  $(x_n)$  generated by algorithm (SFBP) converges strongly to the unique  $u \in S$  as  $n \rightarrow \infty$ .*

Finally, a function  $F : \mathcal{H} \rightarrow \mathbf{R} \cup \{+\infty\}$  is *boundedly inf-compact* if the sets of the form

$$\{x \in \mathcal{H} : \|x\| \leq R, \text{ and } F(x) \leq M\},$$

are relatively compact for every  $R \geq 0$  and  $M \in \mathbf{R}$ .

We shall prove that if  $A$  is the subdifferential of a proper, lower-semicontinuous and convex function  $\Phi_2$ , weak convergence of the sequences generated by the (SFBP) algorithm can be guaranteed if either  $\Phi_2$  is boundedly inf-compact, the penalization parameters satisfy a *subexponential* growth condition, or in the unconstrained case. More precisely, we have the following:

**Theorem 3.** *Let  $(H_0)$  hold with  $A = \partial \Phi_2$ . Assume that any of the following conditions holds:*

- (i)  $\liminf_{n \rightarrow \infty} \lambda_n \beta_n > 0$  and the function  $(\Phi + \Phi_2)$  or  $(\Psi_1 + \Psi_2)$  is boundedly inf-compact;
- (ii)  $\liminf_{n \rightarrow \infty} \lambda_n \beta_n > 0$ ,  $(\lambda_n)$  is bounded and  $\beta_{n+1} - \beta_n \leq K \lambda_{n+1} \beta_{n+1}$  for some  $K > 0$ ; or
- (iii)  $\Psi_1 = \Psi_2 = 0$ .

*Then, the sequence  $(x_n)$  converges weakly to a point in  $S$  as  $n \rightarrow \infty$ . Moreover, convergence is strong in case (i) if  $\Psi_1 + \Psi_2$  is boundedly inf-compact.*

*The sequence  $(x_n)$  minimizes  $\Phi + \Phi_2$  in cases (ii) and (iii):*

$$\lim_{n \rightarrow \infty} (\Phi + \Phi_2)(x_n) = \min_C (\Phi + \Phi_2).$$

The proofs of Theorems 1, 2 and 3 will be completed in Sections 5.2, 5.3 and 5.4, respectively. One cannot expect to have strong convergence in Theorem 3 in general, see the comment after Corollary 6.

**2.3. Inexact computation of the iterates.** Convergence also holds if the iterates are computed inexactly provided the errors are small enough. More precisely, consider the *inexact splitting forward-backward penalty algorithm* given by

$$(SFBP_\varepsilon) \quad \begin{cases} x_1 & \in \mathcal{H}, \\ x_{n+1} & = (I + \lambda_n A + \lambda_n \beta_n \partial \Psi_2)^{-1}(x_n - \lambda_n \nabla \Phi(x_n) - \lambda_n \beta_n \nabla \Psi_1(x_n) - \zeta_n) + \xi_n, \\ & n \geq 1. \end{cases}$$

We recall the following result from [1]:

**Lemma 4.** *Let  $(P_n)$  be a sequence of nonexpansive functions from  $\mathcal{H}$  into  $\mathcal{H}$ . Let  $(\varepsilon_n)$  be a positive sequence in  $\ell^1$ . If every sequence  $(x_n)$  satisfying*

$$x_{n+1} = P_n(x_n), \quad n \geq 1$$

*converges weakly (resp. strongly, resp. weakly or strongly in average), then the same is true for every sequence  $(\tilde{x}_n)$  satisfying*

$$\|\tilde{x}_{n+1} - P_n(\tilde{x}_n)\| \leq \varepsilon_n, \quad n \geq 1.$$

Following the arguments in the proof of [4, Proposition 6.3], we obtain

**Corollary 5.** *Let  $(\zeta_n)$  and  $(\xi_n)$  be nonnegative sequences in  $\ell^1$ , and let  $(x_n)$  verify  $(SFBP_\varepsilon)$ . Then Theorems 1, 2 and 3 remain true.*

**2.4. Forward Backward Backward algorithm and full splitting.** The (SFBP) algorithm is a step toward full splitting. It allows to understand the different roles played by the regular parts -allowing for forward steps- and the general parts -needing backward steps. It requires to compute the resolvent of the sum of two maximal monotone operators, which may be a hard task. The full splitting of the backward step is achieved in [4]. Following the same path, let us define the *splitting forward-backward-backward penalty algorithm* (SFBBP), as follows:

$$(SFBBP) \quad \begin{cases} x_1 & \in \mathcal{H}, \\ x_{n+1} & = (I + \lambda_n \beta_n \partial \Psi_2)^{-1}(I + \lambda_n A)^{-1}(x_n - \lambda_n \nabla \Phi(x_n) - \lambda_n \beta_n \nabla \Psi_1(x_n)), \quad n \geq 1. \end{cases}$$

The complete study of (SFBBP) goes beyond the scope of this paper. We believe that the convergence results should hold, by making use of the techniques in [4].

### 3. COMPARISON WITH CLASSICAL AND MORE RECENT METHODS

In this section we examine some particular cases, where some of the functions or the operator involved in (2) vanish.

**3.1. The backward algorithm.** Taking the two potentials  $\Phi$  and  $\Psi_1$  to be zero, the forward step disappears and the (SFBP) algorithm turns into a purely backward algorithm.

**3.1.1. The unconstrained case : the proximal point algorithm.** If additionally  $\Psi_2$  is zero, we obtain the *proximal point algorithm*:

$$(PROX) \quad \begin{cases} x_1 & \in H, \\ x_{n+1} & = (I + \lambda_n A)^{-1}(x_n) \text{ for all } n \geq 1. \end{cases}$$

This method was originally introduced in [25], using the idea of *proximity operator* from [27]. It was further developed in [35], [14] and [24]. Its popularity is due to the fact that, despite its iteration-complexity, convergence can be granted under minimal hypotheses.

Let  $(x_n)$  be a sequence generated by (PROX) and define  $(z_n)$  and  $(\hat{z}_n)$  and in (5).

**Corollary 6.** *Let  $S \neq \emptyset$  and  $(\lambda_n) \notin \ell^1$ . As  $n \rightarrow \infty$ , we have the following:*



- i) The sequence  $(\widehat{z}_n)$  converges weakly to a point in  $S$ ;
- ii) If  $(\lambda_n) \in \ell^2$ , then the sequence  $(z_n)$  converges weakly to a point in  $S$ ;
- iii) If  $A$  is strongly monotone, then  $(x_n)$  converges strongly to the unique point in  $S$ ; and
- iv) If  $A = \partial\Phi_2$ , then  $(x_n)$  converges weakly to a point in  $S$ , with  $\lim_{n \rightarrow \infty} \Phi_2(x_n) = \min(\Phi_2)$ .

Part i) is [34, Theorem 5.6], part ii) is [24, Theorem II.1.], part iii) is [14, Remark 11] and part iv) is [14, Theorem 9].

A counterexample for strong convergence in case iv) was given in [23], following the ideas in [6]. Therefore, one cannot expect to obtain strong convergence in Theorem 3 in general.

**3.1.2. Penalized algorithms: diagonal proximal algorithm.** In general, we obtain the diagonal proximal algorithm from [4]:

$$(DPA) \quad \begin{cases} x_1 & \in H, \\ x_{n+1} & = (I + \lambda_n A + \lambda_n \beta_n \partial\Psi_2)^{-1}(x_n) \text{ for all } n \geq 1, \end{cases}$$

Hypothesis  $(H_0)$  becomes

$$(H'_0) \quad \begin{cases} i) & \mathbf{T} = A + N_C \text{ is maximal monotone and } S = \mathbf{T}^{-1}(0) \neq \emptyset; \\ ii) & \text{For each } z \in N_C(\mathcal{H}), \sum_{n=1}^{\infty} \lambda_n \beta_n \left[ \Psi_2^*\left(\frac{z}{\beta_n}\right) - \sigma_C\left(\frac{z}{\beta_n}\right) \right] < \infty; \\ iii) & \sum_{n=1}^{\infty} \lambda_n = +\infty. \end{cases}$$

Let  $(x_n)$  be a sequence generated by (DPA) and define  $(z_n)$  and  $(\widehat{z}_n)$  and in (5).

**Corollary 7.** *Let  $(H'_0)$  hold. As  $n \rightarrow \infty$ , we have the following:*

- i) The sequence  $(\widehat{z}_n)$  converges weakly to a point in  $S$ ;
- ii) If  $(\lambda_n) \in \ell^2$ , then the sequence  $(z_n)$  converges weakly to a point in  $S$ ; and
- iii) If  $A$  is strongly monotone, then  $(x_n)$  converges strongly to the unique point in  $S$ .

Part i) is [4, Theorem 3.3] and part iii) is [4, Theorem 3.4]. For the weak convergence, we have

**Corollary 8.** *Let  $(H'_0)$  hold with  $A = \partial\Phi_2$ . Assume any of the following conditions holds:*

- (i)  $\liminf_{n \rightarrow \infty} \lambda_n \beta_n > 0$  and either  $\Phi_2$  or  $\Psi_2$  is boundedly inf-compact; or
- (ii)  $\liminf_{n \rightarrow \infty} \lambda_n \beta_n > 0$ ,  $(\lambda_n)$  is bounded and  $\beta_{n+1} - \beta_n \leq K \lambda_{n+1} \beta_{n+1}$  for some  $K > 0$ .

*Then, the sequence  $(x_n)$  converges weakly to a point in  $S$  as  $n \rightarrow \infty$ , with*

$$\lim_{n \rightarrow \infty} \Phi_2(x_n) = \min_C(\Phi_2).$$

*Moreover, convergence is strong in case (i) if  $\Psi_2$  is boundedly inf-compact.*

The hypotheses in part ii) are very close to, but slightly different from those corresponding to cases ii) and iii) of [4, Theorem 3.6].

**3.2. The forward algorithm.** Taking the operator  $A$  and the potential  $\Psi_2$  to be zero, the backward step disappears and the (SFBP) algorithm turns into a purely forward algorithm.

*The gradient method.* If additionally  $\Psi_1 = 0$ , we obtain the *gradient method*, which dates back to [20]:

$$(GRAD) \quad \begin{cases} x_1 & \in H, \\ x_{n+1} & = x_n - \lambda_n \nabla \Phi(x_n) \text{ for all } n \geq 1. \end{cases}$$

Let  $(x_n)$  be a sequence generated by (GRAD).

**Corollary 9.** *Let  $\Phi$  be a convex function with Lipschitz-continuous gradient. Assume  $S \neq \emptyset$  and  $(\lambda_n) \in \ell^2 \setminus \ell^1$ . As  $n \rightarrow \infty$ , the sequence  $(x_n)$  converges weakly to a point in  $S$ , with  $\lim_{n \rightarrow \infty} \Phi(x_n) = \min(\Phi)$ .*

This is not the most general convergence result for the gradient method. The hypothesis  $(\lambda_n) \in \ell^2$  may be replaced by  $\limsup_{n \rightarrow \infty} \lambda_n < 2/L_\Phi$  (see [32, Teorema 9.6]). Intermediate results are proved in [11, Paragraph 1.2.13], assuming the step sizes tend to zero; and in [17, Theorem 3], under a very precise condition on the step sizes:  $\delta_1 \leq \lambda_n \leq \frac{2}{L_\Phi}(1 - \delta_2)$  with  $\delta_1, \delta_2 > 0$  such that  $\frac{L_\Phi}{2}\delta_1 + \delta_2 < 1$ . The last two are proved in  $\mathcal{H} = \mathbf{R}^n$ , but the proof can be easily adapted to the Hilbert-space framework.

**3.2.1. Penalized algorithm: a diagonal gradient scheme.** If  $A \equiv 0$  we obtain the *diagonal gradient scheme* studied in [33], namely:

$$(DGS) \quad \begin{cases} x_1 & \in \mathcal{H}, \\ x_{n+1} & = x_n - \lambda_n \nabla \Phi(x_n) - \lambda_n \beta_n \nabla \Psi_1(x_n), \quad n \geq 1. \end{cases}$$

Hypothesis  $(H_0)$  becomes

$$(H_0''') \quad \begin{cases} i) & S = \mathbf{T}^{-1}(0) \neq \emptyset; \\ ii) & \nabla \Phi \text{ is } L_\Phi\text{-Lipschitz-continuous and } \nabla \Psi_1 \text{ is } L_{\Psi_1}\text{-Lipschitz-continuous}; \\ iii) & \text{For each } z \in N_C(\mathcal{H}), \quad \sum_{n=1}^{\infty} \lambda_n \beta_n \left[ \Psi_1^*\left(\frac{z}{\beta_n}\right) - \sigma_C\left(\frac{z}{\beta_n}\right) \right] < \infty; \\ iv) & \sum_{n=1}^{\infty} \lambda_n = +\infty, \quad \sum_{n=1}^{\infty} \lambda_n^2 < +\infty, \quad \text{and} \quad \limsup_{n \rightarrow \infty} (L_{\Psi_1} \lambda_n \beta_n) < 2. \end{cases}$$

Then we have:

**Corollary 10.** *Let  $(H_0''')$  hold, and assume that any of the following conditions holds:*

- (i)  $\liminf_{n \rightarrow \infty} \lambda_n \beta_n > 0$  and the function  $\Phi$  or  $\Psi_1$  is boundedly inf-compact; or
- (ii)  $\liminf_{n \rightarrow \infty} \lambda_n \beta_n > 0$ ,  $(\lambda_n)$  is bounded and  $\beta_{n+1} - \beta_n \leq K \lambda_{n+1} \beta_{n+1}$  for some  $K > 0$ .

*Then, the sequence  $(x_n)$  converges weakly to a point in  $S$  as  $n \rightarrow \infty$ , with*

$$\lim_{n \rightarrow \infty} \Phi(x_n) = \min_C(\Phi).$$

*Moreover, convergence is strong in case (i) if  $\Psi_1$  is boundedly inf-compact.*

Part ii) was given in [33, Theorem 2.1] with slightly different hypotheses.

### 3.3. The forward-backward splitting.

**3.3.1. with no penalization.** If  $A \neq 0$  and  $\Phi \neq 0$  we obtain the forward-backward splitting:

$$(FB) \quad \begin{cases} x_1 & \in H, \\ x_{n+1} & = (I + \lambda_n A)^{-1}(x_n - \lambda_n \nabla \Phi(x_n)) \quad \text{for all } n \geq 1. \end{cases}$$

This method combines the previous two, inheriting their virtues and drawbacks. A particularly interesting case is the *projected gradient method* described as follows: Let  $C$  be a nonempty, closed and convex subset of  $\mathcal{H}$  and set  $A = N_C$ . Then (FB) becomes

$$(PG) \quad \begin{cases} x_1 & \in H, \\ x_{n+1} & = \text{Proj}_C(x_n - \lambda_n \nabla \Phi(x_n)) \quad \text{for all } n \geq 1. \end{cases}$$

This is useful for minimization problems of the form

$$\min\{\Phi(x) : x \in C\},$$

when the projection onto the set  $C$  is easily performed.

Let  $(x_n)$  be a sequence generated by (FB).

**Corollary 11.** *Let  $\Phi$  be a convex function with Lipschitz-continuous gradient. Assume  $S \neq \emptyset$  and  $(\lambda_n) \in \ell^2 \setminus \ell^1$ . As  $n \rightarrow \infty$ , we have the following:*

- i) *The sequence  $(\hat{z}_n)$  converges weakly to a point in  $S$ ;*
- ii) *If  $(\lambda_n) \in \ell^2$ , then the sequence  $(z_n)$  converges weakly to a point in  $S$ ;*
- iii) *If  $A$  is strongly monotone, then  $(x_n)$  converges strongly to the unique point in  $S$ ; and*
- iv) *If  $A = \partial\Phi_2$ , then  $(x_n)$  converges weakly to a point in  $S$ , with  $\lim_{n \rightarrow \infty} \Phi_2(x_n) = \min(\Phi_2)$ .*

The results in [16, Theorem 1] and [31, Theorem 2] are closely related to part ii). Although they consider a maximal monotone operator  $B$  instead of  $\nabla\Phi$  (which is a more general framework), their results rely on a  $\ell^2$ -summability condition – that is difficult to check in practice – concerning a sequence  $(w_n)$  satisfying  $w_n \in \lambda_n B(x_n)$ . Analogously, [31, Corollary 1] is close to part iv).

If the step sizes are bounded from below by a positive constant, then the sequence  $(x_n)$  converges weakly, even when  $A$  is a maximal monotone operator and  $\nabla\Phi$  is replaced by a *cocoercive* function  $B$  (see [21, Corollary 6.5]). A function  $B : \mathcal{H} \rightarrow \mathcal{H}$  is cocoercive if  $\langle Bx - By, x - y \rangle \geq \beta \|Bx - By\|^2$  for all  $x, y \in \mathcal{H}$ .

*Smooth penalization.* If  $\Phi \equiv 0$ , we obtain the forward-backward-penalty scheme studied in [5] and [29]:

$$(FBP) \quad \begin{cases} x_1 & \in \mathcal{H}, \\ x_{n+1} & = (I + \lambda_n A)^{-1}(x_n - \lambda_n \beta_n \nabla \Psi_1(x_n)), \quad n \geq 1. \end{cases}$$

Hypothesis  $(H_0)$  becomes

$$(H_0'') \quad \begin{cases} i) & \mathbf{T} = A + N_C \text{ is maximal monotone and } S = \mathbf{T}^{-1}(0) \neq \emptyset; \\ ii) & \nabla \Psi_1 \text{ is } L_{\Psi_1}\text{-Lipschitz-continuous;} \\ iii) & \text{For each } z \in N_C(\mathcal{H}), \quad \sum_{n=1}^{\infty} \lambda_n \beta_n \left[ \Psi_1^*\left(\frac{z}{\beta_n}\right) - \sigma_C\left(\frac{z}{\beta_n}\right) \right] < \infty; \\ iv) & \sum_{n=1}^{\infty} \lambda_n = +\infty, \quad \text{and} \quad \limsup_{n \rightarrow \infty} (L_{\Psi_1} \lambda_n \beta_n) < 2. \end{cases}$$

**Corollary 12.** *Assume that  $(H_0'')$  holds. As  $n \rightarrow \infty$ , we have the following:*

- i) *The sequence  $(\hat{z}_n)$  converges weakly to a point in  $S$ ;*
- ii) *If  $(\lambda_n) \in \ell^2$ , then the sequence  $(z_n)$  converges weakly to a point in  $S$ ;*
- iii) *If  $A$  is strongly monotone, then  $(x_n)$  converges strongly to the unique point in  $S$ .*

Parts ii) and iii) yield [5, Theorem 12]. Observe that, in part i), convergence is proved without the  $\ell^2$ -summability assumption, unlike in [5]. For the weak convergence we have the following:

**Corollary 13.** *Let  $(H_0'')$  hold with  $A = \partial\Phi_2$ . Assume that any of the following conditions holds:*

- (i)  $\liminf_{n \rightarrow \infty} \lambda_n \beta_n > 0$  and the function  $\Phi_2$  or  $\Psi_1$  is boundedly inf-compact; or
- (ii)  $\liminf_{n \rightarrow \infty} \lambda_n \beta_n > 0$ ,  $(\lambda_n)$  is bounded and  $\beta_{n+1} - \beta_n \leq K \lambda_{n+1} \beta_{n+1}$  for some  $K > 0$ .

*Then, the sequence  $(x_n)$  converges weakly to a point in  $S$  as  $n \rightarrow \infty$ , with*

$$\lim_{n \rightarrow \infty} \Phi_2(x_n) = \min_C(\Phi_2).$$

*Moreover, convergence is strong in case (i) if  $\Psi_1$  is boundedly inf-compact.*

Part i) yields [5, Theorem 16], Part ii) yields [29, Theorem 1]. Both cited results additionally assume the  $\ell^2$ -summability of the step sizes.

**Remark 14.** *The main results of this paper appeared in the PhD thesis of Nahla Noun [28]. Simultaneously, Boř and Csetnek [12] extended the forward-backward results of [5] in order to solve the variational inequality*

$$0 \in Ax + Dx + N_C(x),$$

where  $A$  is a maximal monotone operator,  $D$  a cocoercive operator, and  $C$  is the set of zeroes of another maximal monotone operator. Their framework is related but different from ours and cannot be immediately compared.

#### 4. AN ILLUSTRATION: SPARSE-OPTIMAL CONTROL OF A LINEAR SYSTEM OF ODE'S

Given  $y_0 \in \mathbf{R}^n$ ,  $A : [0, T] \rightarrow \mathbf{R}^{n \times n}$ ,  $B : [0, T] \rightarrow \mathbf{R}^{n \times m}$ , and  $c : [0, T] \rightarrow \mathbf{R}^n$ , consider the control system

$$(CS) \quad \begin{cases} \dot{y}(t) &= A(t)y(t) + B(t)u(t) + c(t), & t \in (0, T) \\ y(0) &= y_0. \end{cases}$$

We assume that the functions  $A$ ,  $B$  and  $c$  are bounded and sufficiently regular so that, for each  $u \in L^\infty(0, T; \mathbf{R}^m)$ , the system (CS) has a unique solution  $y_u : [0, T] \rightarrow \mathbf{R}^n$ , which is an absolutely continuous function such that  $y_u(0) = y_0$  and the differential equation holds almost everywhere.

We are interested in the *optimal control problem*

$$(OCP) \quad \min \left\{ \frac{1}{2} \|y_u - \bar{y}\|_{L^2(0, T; \mathbf{R}^n)}^2 + \|u\|_{L^2(0, T; \mathbf{R}^m)}^2 + \|u\|_{L^1(0, T; \mathbf{R}^m)} : u \in \mathcal{U} \right\},$$

where  $\bar{y}$  is a *reference* trajectory, and the set of admissible controls is

$$\mathcal{U} = \{ u : [0, T] \rightarrow \mathbf{R}^m : u \text{ is measurable and } |u_i(t)| \leq 1 \text{ a.e. for each } i = 1, \dots, m \}.$$

The term  $\|u\|_{L^2(0, T; \mathbf{R}^m)}^2$  can be interpreted as a measure of the energy invested in controlling the system, and the minimization of the term  $\|u\|_{L^1(0, T; \mathbf{R}^m)}$  is known to induce sparsity of the solution.

Let  $R : [0, T] \rightarrow \mathbf{R}^{n \times n}$  be the resolvent of the matrix equation  $\dot{X} = AX$  with initial condition  $X(0) = I$ . The pair  $(u, y)$  satisfies (CS) if, and only if,

$$y(t) = R(t)y_0 + R(t) \int_0^t R(s)^{-1} [B(s)u(s) + c(s)] dt.$$

This, in turn, is equivalent to

$$\mathcal{M}(u, y) + z_0 = 0,$$

where we have written

$$\mathcal{M}(u, y)(t) = -y(t) + R(t) \int_0^t R(s)^{-1} B(s)u(s) dt, \quad \text{and} \quad z_0(t) = R(t)y_0 + R(t) \int_0^t R(s)^{-1} c(s) dt.$$

Set  $H = L^2(0, T; \mathbf{R}^m) \times L^2(0, T; \mathbf{R}^n)$ . Since  $\mathcal{M}$  is a bounded linear operator from  $H$  to  $L^2(0, T; \mathbf{R}^n)$ , the function  $\Psi_1 : H \rightarrow \mathbf{R}$ , defined by

$$\Psi_1(u, y) = \frac{1}{2} \|\mathcal{M}(u, y) + z_0\|_{L^2(0, T; \mathbf{R}^n)}^2,$$

is convex and continuously differentiable. On the other hand, since  $\mathcal{U}$  is nonempty, closed in  $L^2(0, T; \mathbf{R}^m)$  and convex, the function  $\Psi_2 : H \rightarrow \mathbf{R} \cup \{+\infty\}$ , defined by

$$\Psi_2(u, y) = \delta_{\mathcal{U}}(u)$$

(the indicator function of the set  $\mathcal{U}$ ), is proper, lower-semicontinuous and convex. Moreover, the pair  $(u, y)$  satisfies (CS) with  $u \in \mathcal{U}$  if, and only if,  $(u, y) \in \text{Argmin}(\Psi_1 + \Psi_2)$ . With this notation, the optimal control problem (OCP) is equivalent to the *constrained optimization problem*

$$(\text{COP}) \quad \min \{ \Phi_1(u, y) + \Phi_2(u, y) : (u, y) \in \text{Argmin}(\Psi_1 + \Psi_2) \},$$

where  $\Phi_1 : H \rightarrow \mathbf{R}$  is the convex and continuously differentiable function defined by

$$\Phi_1(u, y) = \frac{1}{2} \|y - \bar{y}\|_{L^2(0, T; \mathbf{R}^n)}^2 + \|u\|_{L^2(0, T; \mathbf{R}^m)}^2,$$

and  $\Phi_2 : H \rightarrow \mathbf{R} \cup \{+\infty\}$  is the proper, lower-semicontinuous and convex function given by

$$\Phi_2(u, y) = \|u\|_{L^1(0, T; \mathbf{R}^m)}.$$

## 5. PROOFS

Recall that, by assumption  $(H_0)$ , the monotone operator  $\mathbf{T} = A + \nabla\Phi + N_C$  is maximal and  $S = \mathbf{T}^{-1}(0) \neq \emptyset$ . Its domain is  $\text{dom}(\mathbf{T}) = C \cap \text{dom}(A)$ . The functions  $\Phi$  and  $\Psi_1$  are differentiable and their gradients  $\nabla\Phi$  and  $\nabla\Psi_1$  are Lipschitz-continuous with constants  $L_\Phi$  and  $L_{\Psi_1}$ , respectively. Since  $\min \Psi_1 = \min \Psi_2 = 0$ , the function  $\Psi_1 + \Psi_2$  vanishes on  $C = \text{Argmin}(\Psi_1) \cap \text{Argmin}(\Psi_2)$ .

The proofs of Theorems 1–3 ultimately rely on a well-known tool from [30] (see the proper statement in [7]) and [31] which gives weak convergence without a priori knowledge of the limit.

**Lemma 15.** *Given a sequence  $(x_n)$  in  $\mathcal{H}$ , a sequence  $(\lambda_n) \notin \ell^1$  of positive numbers, set*

$$z_n = \frac{1}{\tau_n} \sum_{k=1}^n \lambda_k x_k \quad \text{and} \quad \hat{z}_n = \frac{1}{\tau_n} \sum_{k=1}^n \lambda_k x_{k+1}, \quad \text{with} \quad \tau_n = \sum_{k=1}^n \lambda_k.$$

*Let  $S$  be a subset of  $\mathcal{H}$  and assume that*

(i) *for every  $x \in S$ ,  $\lim_{n \rightarrow \infty} \|x_n - x\|$  exists;*

(ii) *every weak cluster point of  $(x_n)$ , respectively  $(z_n)$ , resp.  $(\hat{z}_n)$ , lies in  $S$ .*

*Then  $(x_n)$ , respectively  $(z_n)$ , resp.  $(\hat{z}_n)$ , converges weakly to a point in  $S$  as  $n \rightarrow \infty$ .*

The core of the convergence analysis is the following estimation:

**Lemma 16.** *There exist  $a, b, c, d, e > 0$  such that, for every  $u \in \text{dom}(\mathbf{T})$ ,  $z \in Au$ ,  $v \in N_C(u)$ , and  $w = z + \nabla\Phi(u) + v$ , the following inequality holds for  $n$  large enough*

$$\begin{aligned} (6) \quad & [1 - aL_\Phi\lambda_n^2] \|x_{n+1} - u\|^2 - \|x_n - u\|^2 + b\|x_{n+1} - x_n\|^2 + c\lambda_n \|\nabla\Phi(x_{n+1}) - \nabla\Phi(u)\|^2 \\ & + \frac{d}{2} \lambda_n \beta_n (\Psi_1 + \Psi_2)(x_{n+1}) + e\lambda_n \beta_n \|\nabla\Psi_1(x_n)\|^2 \\ & \leq \frac{d}{2} \lambda_n \beta_n \left[ (\Psi_1 + \Psi_2)^*\left(\frac{4v}{d\beta_n}\right) - \sigma_C\left(\frac{4v}{d\beta_n}\right) \right] + 2\lambda_n \langle w, u - x_{n+1} \rangle. \end{aligned}$$

The proof uses standard convex analysis tools along with very careful estimations. Since it is highly technical, it will be given below in Subsection 5.1. A straightforward consequence of Lemma 16 is the following proposition, which contains the basic properties of the algorithm, including the first assumption of Lemma 15:

**Proposition 17.** *Assume  $(H_0)$ ,  $(\lambda_n L_\Phi) \in \ell^2$ , and let  $(x_n)$  be a sequence generated by the (SFBP) Algorithm. Then the following holds:*

i) *For every  $u \in S$ ,  $\lim_{n \rightarrow \infty} \|x_n - u\|$  exists.*

- ii) The series  $\sum_{n \geq 1} \|x_{n+1} - x_n\|^2$ ,  $\sum_{n \geq 1} \lambda_n \beta_n \Psi_1(x_n)$ ,  $\sum_{n \geq 1} \lambda_n \beta_n \Psi_2(x_n)$ ,  $\sum_{n \geq 1} \lambda_n \|\nabla \Phi(x_{n+1}) - \nabla \Phi(u)\|^2$  and  $\sum_{n \geq 1} \lambda_n \beta_n \|\nabla \Psi_1(x_n)\|^2$  converge.
- iii) If moreover  $\liminf_{n \rightarrow \infty} \lambda_n \beta_n > 0$ , then  $\sum_{n \geq 1} \Psi_1(x_n)$  and  $\sum_{n \geq 1} \Psi_2(x_n)$  converge,  $\lim_{n \rightarrow \infty} \Psi_1(x_n) = \lim_{n \rightarrow \infty} \Psi_2(x_n) = 0$ , and every weak cluster point of the sequence  $(x_n)$  lies in  $C$ .

Indeed, if  $u \in S$ , then  $0 \in Au + \nabla \Phi(u) + N_C(u)$ . Write  $0 = z + \nabla \Phi(u) + v$  with  $z \in Au$  and  $v \in N_C(u)$ . For  $n$  large enough, Lemma 16 gives

$$[1 - aL_\Phi \lambda_n^2] \|x_{n+1} - u\|^2 - \|x_n - u\|^2 + b \|x_{n+1} - x_n\|^2 + c \lambda_n \|\nabla \Phi(x_{n+1}) - \nabla \Phi(u)\|^2 + \frac{d}{2} \lambda_n \beta_n (\Psi_1 + \Psi_2)(x_{n+1}) + e \lambda_n \beta_n \|\nabla \Psi_1(x_n)\|^2 \leq \frac{d}{2} \lambda_n \beta_n \left[ (\Psi_1 + \Psi_2)^* \left( \frac{4v}{d\beta_n} \right) - \sigma_C \left( \frac{4v}{d\beta_n} \right) \right]$$

Since the right-hand side is summable, all the parts of Proposition 17 ensue from the following elementary fact concerning the convergence of real sequences:

**Lemma 18.** *Let  $(a_n)$ ,  $(\delta_n)$  and  $(\varepsilon_n)$  be nonnegative and let  $(\xi_n)$  be bounded from below. Assume*

$$(1 - a_n)\xi_{n+1} - \xi_n + \delta_n \leq \varepsilon_n$$

*for all  $n$  large enough. If  $(a_n)$  and  $(\varepsilon_n)$  belong to  $\ell^1$ , then  $(\xi_n)$  is convergent and  $(\delta_n)$  belongs to  $\ell^1$ .*

**5.1. Proof of Lemma 16.** Take  $u \in \text{dom}(\mathbf{T})$ ,  $z \in Au$ ,  $v \in N_C(u)$ , and  $w = z + \nabla \Phi(u) + v$ . Rewrite algorithm (SFBP) as

$$(7) \quad \frac{x_n - x_{n+1}}{\lambda_n} - \nabla \Phi(x_n) - v_{n+1} - \beta_n w_{n+1} - \beta_n \nabla \Psi_1(x_n) = 0.$$

with  $v_{n+1} \in Ax_{n+1}$  and  $w_{n+1} \in \partial \Psi_2(x_{n+1})$ .

**Claim 19.** *The following inequality holds for every  $n$ :*

$$(8) \quad \|x_{n+1} - u\|^2 - \|x_n - u\|^2 + \|x_{n+1} - x_n\|^2 + 2\lambda_n \langle \nabla \Phi(x_n) - \nabla \Phi(u), x_{n+1} - u \rangle + 2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_{n+1} - u \rangle \leq 2\lambda_n \langle \nabla \Phi(u) + z, u - x_{n+1} \rangle - 2\lambda_n \beta_n \Psi_2(x_{n+1}).$$

**Proof.** The monotonicity of  $A$  at points  $u$  and  $x_{n+1}$  gives

$$(9) \quad 2\lambda_n \langle v_{n+1}, u - x_{n+1} \rangle \leq 2\lambda_n \langle z, u - x_{n+1} \rangle.$$

The subdifferential inequality for function  $\Psi_2$  at  $x_{n+1}$  with  $w_{n+1} \in \partial \Psi_2(x_{n+1})$  gives

$$(10) \quad 2\lambda_n \beta_n \langle w_{n+1}, u - x_{n+1} \rangle \leq -2\lambda_n \beta_n \Psi_2(x_{n+1}).$$

To conclude, it suffices to sum (9) and (10), use (7), along with the fact that

$$2\langle x_{n+1} - x_n, x_{n+1} - u \rangle = \|x_{n+1} - u\|^2 - \|x_n - u\|^2 + \|x_{n+1} - x_n\|^2,$$

and rearrange the terms conveniently. ■

The next step is to estimate some of the terms in (8) (observe that the last two terms on the left-hand side vanish if  $\nabla \Phi$  is constant or  $\Psi_1 \equiv 0$ , which correspond to the cases  $L_\Phi = 0$  and  $L_{\Psi_1} = 0$ , respectively). At different points we shall use the Baillon-Haddad Theorem [10]:

**Lemma 20** (Baillon-Haddad Theorem). *Let  $f : \mathcal{H} \rightarrow \mathbf{R}$  be a convex differentiable function and  $L_f > 0$ . Then  $\nabla f$  is Lipschitz continuous with constant  $L_f$  if and only if  $\nabla f$  is  $\frac{1}{L_f}$ -cocoercive.*

We also use the following Descent Lemma (see, for example [11]):



**Lemma 21** (Descent Lemma). *Let  $f : \mathcal{H} \rightarrow \mathbf{R}$  be continuously differentiable such that  $\nabla f$  is Lipschitz continuous with constant  $L_f$ . Then, for every  $x$  and  $y$  in  $\mathcal{H}$ ,*

$$f(x + y) \leq f(x) + \langle \nabla f(x), y \rangle + \frac{L_f}{2} \|y\|^2.$$

We have the following:

**Claim 22.** *Assume  $\nabla \Phi$  is not constant. For every  $\eta > 0$  we have:*

$$2\lambda_n \langle \nabla \Phi(x_n) - \nabla \Phi(u), x_{n+1} - u \rangle \geq \frac{-\lambda_n^2}{\eta} \|x_{n+1} - u\|^2 - \eta L_\Phi^2 \|x_{n+1} - x_n\|^2 + \frac{2\lambda_n}{L_\Phi} \|\nabla \Phi(x_{n+1}) - \nabla \Phi(u)\|^2.$$

**Proof.** Write

$$\langle \nabla \Phi(x_n) - \nabla \Phi(u), x_{n+1} - u \rangle = \langle \nabla \Phi(x_n) - \nabla \Phi(x_{n+1}), x_{n+1} - u \rangle + \langle \nabla \Phi(x_{n+1}) - \nabla \Phi(u), x_{n+1} - u \rangle.$$

We easily see that

$$2\lambda_n \langle \nabla \Phi(x_n) - \nabla \Phi(x_{n+1}), x_{n+1} - u \rangle \geq \frac{-1}{\eta} \lambda_n^2 \|x_{n+1} - u\|^2 - \eta L_\Phi^2 \|x_{n+1} - x_n\|^2.$$

On the other hand, Lemma 20 implies

$$2\lambda_n \langle \nabla \Phi(x_{n+1}) - \nabla \Phi(u), x_{n+1} - u \rangle \geq \frac{2\lambda_n}{L_\Phi} \|\nabla \Phi(x_{n+1}) - \nabla \Phi(u)\|^2.$$

The result follows immediately. ■

**Claim 23.** *Assume  $\Psi_1 \neq 0$ . For all  $\eta, \theta > 0$  and  $n \in \mathbf{N}$  we have*

$$\begin{aligned} 2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_{n+1} - u \rangle &\geq \left[ \frac{2}{(1+\eta)L_{\Psi_1}} - \frac{(1+\theta)}{1+\eta} \lambda_n \beta_n \right] \lambda_n \beta_n \|\nabla \Psi_1(x_n)\|^2 \\ &\quad + \frac{2\eta}{1+\eta} \lambda_n \beta_n \Psi_1(x_{n+1}) - \left[ \frac{1}{(1+\theta)(1+\eta)} + \frac{\eta L_{\Psi_1}}{1+\eta} \lambda_n \beta_n \right] \|x_{n+1} - x_n\|^2. \end{aligned}$$

**Proof.** Write

$$(11) \quad 2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_{n+1} - u \rangle = 2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_{n+1} - x_n \rangle + 2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_n - u \rangle.$$

A convex combination of the bounds given by Lemma 20 and the subdifferential inequality gives

$$(12) \quad \langle \nabla \Psi_1(x_n), x_n - u \rangle \geq \frac{1}{(1+\eta)L_{\Psi_1}} \|\nabla \Psi_1(x_n)\|^2 + \frac{\eta}{1+\eta} \Psi_1(x_n)$$

for any  $\eta > 0$ . Now take  $\theta > 0$  and use the identity

$$\begin{aligned} \frac{1}{1+\theta} \|x_{n+1} - x_n + (1+\theta)\lambda_n \beta_n \nabla \Psi_1(x_n)\|^2 &= \\ \frac{1}{1+\theta} \|x_{n+1} - x_n\|^2 + (1+\theta)\lambda_n^2 \beta_n^2 \|\nabla \Psi_1(x_n)\|^2 &+ 2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_{n+1} - x_n \rangle, \end{aligned}$$

to obtain

$$2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_{n+1} - x_n \rangle \geq -\frac{1}{1+\theta} \|x_{n+1} - x_n\|^2 - (1+\theta)\lambda_n^2 \beta_n^2 \|\nabla \Psi_1(x_n)\|^2.$$

On the other hand, Lemma 21 at  $x_n$  and  $x_{n+1}$  gives

$$2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_{n+1} - x_n \rangle \geq 2\lambda_n \beta_n [\Psi_1(x_{n+1}) - \Psi_1(x_n)] - L_{\Psi_1} \lambda_n \beta_n \|x_{n+1} - x_n\|^2.$$

A convex combination of the last two inequalities produces

$$(13) \quad 2\lambda_n\beta_n\langle\nabla\Psi_1(x_n), x_{n+1} - x_n\rangle \geq -\frac{1+\theta}{1+\eta}\lambda_n^2\beta_n^2\|\nabla\Psi_1(x_n)\|^2 \\ + \frac{2\eta}{1+\eta}\lambda_n\beta_n[\Psi_1(x_{n+1}) - \Psi_1(x_n)] - \left[\frac{1}{(1+\eta)(1+\theta)} + \frac{\eta L_{\Psi_1}}{1+\eta}\lambda_n\beta_n\right]\|x_{n+1} - x_n\|^2.$$

Finally, use (12) and (13) in (11) to conclude.  $\blacksquare$

**Claim 24.** *There exist  $a, b, c, d, e > 0$  such that for all sufficiently large  $n \in \mathbf{N}$  we have*

$$(14) \quad (1 - aL_\Phi\lambda_n^2)\|x_{n+1} - u\|^2 - \|x_n - u\|^2 + b\|x_{n+1} - x_n\|^2 + c\lambda_n\|\nabla\Phi(x_{n+1}) - \nabla\Phi(u)\|^2 \\ + d\lambda_n\beta_n(\Psi_1 + \Psi_2)(x_{n+1}) + e\lambda_n\beta_n\|\nabla\Psi_1(x_n)\|^2 \leq 2\lambda_n\langle\nabla\Phi(u) + z, u - x_{n+1}\rangle.$$

**Proof.** We focus on the case where  $\nabla\Phi$  is not constant and  $\Psi_1 \not\equiv 0$ . The other cases are simpler and left to the reader. Claims 19, 22 and 23, and the fact that

$$-2\lambda_n\beta_n\Psi_2(x_{n+1}) \leq \frac{-2\eta}{1+\eta}\lambda_n\beta_n\Psi_2(x_{n+1})$$

for every  $\eta > 0$ , together imply

$$\left[1 - \frac{\lambda_n^2}{\eta}\right]\|x_{n+1} - u\|^2 - \|x_n - u\|^2 + \left[1 - \frac{1}{(1+\theta)(1+\eta)} - \frac{\eta L_{\Psi_1}}{1+\eta}\lambda_n\beta_n - \eta L_\Phi^2\right]\|x_{n+1} - x_n\|^2 \\ + \frac{2}{L_\Phi}\lambda_n\|\nabla\Phi(x_{n+1}) - \nabla\Phi(u)\|^2 + \left[\frac{2}{(1+\eta)L_{\Psi_1}} - \frac{(1+\theta)}{1+\eta}\lambda_n\beta_n\right]\lambda_n\beta_n\|\nabla\Psi_1(x_n)\|^2 \\ + \frac{2\eta}{1+\eta}\lambda_n\beta_n(\Psi_1 + \Psi_2)(x_{n+1}) \leq 2\lambda_n\langle\nabla\Phi(u) + z, u - x_{n+1}\rangle$$

for every  $\eta, \theta > 0$ . Set  $\Gamma = \limsup_{n \rightarrow \infty} L_{\Psi_1}\lambda_n\beta_n < 2$  and take  $\theta_0 > 0$  small enough such that

$$2 - (1 + \theta_0)\lambda_n\beta_n L_{\Psi_1} \geq 2 - (1 + \theta_0)\Gamma > 0$$

for all sufficiently large  $n$ . Since

$$\lim_{\eta \rightarrow 0^+} \left[1 - \frac{1}{(1+\theta)(1+\eta)} - \frac{\eta\Gamma L_{\Psi_1}}{1+\eta} - \eta L_\Phi^2\right] = 1 - \frac{1}{(1+\theta)} > 0,$$

we can take  $\eta_0 > 0$  so that (14) holds with  $a = \frac{1}{\eta_0 L_\Phi}$ ,  $b = 1 - \frac{1}{(1+\theta_0)(1+\eta_0)} - \frac{\eta_0\Gamma L_{\Psi_1}}{1+\eta_0} - \eta_0 L_\Phi^2$ ,  $c = \frac{2}{L_\Phi}$ ,  $d = \frac{2\eta_0}{1+\eta_0}$ , and  $e = \frac{2(1+\eta_0)(1+\theta_0)L_{\Psi_1}\Gamma}{(1+\eta_0)L_{\Psi_1}}$ , which are all positive.  $\blacksquare$

**Proof of Lemma 16, completed.** Observe that

$$2\lambda_n\langle\nabla\Phi(u) + z, u - x_{n+1}\rangle - \frac{d}{2}\lambda_n\beta_n(\Psi_1 + \Psi_2)(x_{n+1}) \\ = 2\lambda_n\langle v, x_{n+1} - u\rangle - \frac{d}{2}\lambda_n\beta_n(\Psi_1 + \Psi_2)(x_{n+1}) - 2\lambda_n\langle w, x_{n+1} - u\rangle \\ = \frac{d}{2}\lambda_n\beta_n \left[ \left\langle \frac{4v}{d\beta_n}, x_{n+1} \right\rangle - (\Psi_1 + \Psi_2)(x_{n+1}) - \left\langle \frac{4v}{d\beta_n}, u \right\rangle \right] - 2\lambda_n\langle w, x_{n+1} - u\rangle \\ \leq \frac{d}{2}\lambda_n\beta_n \left[ (\Psi_1 + \Psi_2)^* \left( \frac{4v}{d\beta_n} \right) - \sigma_C \left( \frac{4v}{d\beta_n} \right) \right] - 2\lambda_n\langle w, x_{n+1} - u\rangle$$

because  $\frac{4v}{d\beta_n} \in N_C(u)$  implies  $\sigma_C\left(\frac{4v}{d\beta_n}\right) = \langle \frac{4v}{d\beta_n}, u \rangle$ . Whence

$$(15) \quad 2\lambda_n \langle p - z, x_{n+1} - u \rangle \leq \frac{d}{2} \lambda_n \beta_n (\Psi_1 + \Psi_2)(x_{n+1}) \\ + \frac{d}{2} \lambda_n \beta_n \left[ (\Psi_1 + \Psi_2)^* \left( \frac{4v}{d\beta_n} \right) - \sigma_C \left( \frac{4v}{d\beta_n} \right) \right] + 2\lambda_n \langle w, u - x_{n+1} \rangle.$$

We obtain (6) by using (15) in (14) and rearranging the terms containing  $(\Psi_1 + \Psi_2)(x_{n+1})$ .  $\blacksquare$

**5.2. Weak ergodic convergence: proof of Theorem 1.** In view of Lemma 15 and part i) of Proposition 17, it suffices to prove that every weak cluster point of the sequence  $(z_n)$ , respectively  $(\widehat{z}_n)$ , lies in  $S$ . By maximal monotonicity of  $\mathbf{T}$ , a point  $\bar{x}$  belongs to  $S$  if and only if  $\langle w, u - \bar{x} \rangle \geq 0$  for all  $u \in C \cap \text{dom}(A)$  and all  $w \in \mathbf{T}(u)$ .

We begin with  $(\widehat{z}_n)$ . Take any  $u \in C \cap \text{dom}(A)$  and  $w \in \mathbf{T}(u)$ . By Lemma 16, we have

$$(16) \quad \|x_{n+1} - u\|^2 - \|x_n - u\|^2 \leq \frac{d}{2} \lambda_n \beta_n \left[ (\Psi_1 + \Psi_2)^* \left( \frac{4v}{d\beta_n} \right) - \sigma_C \left( \frac{4v}{d\beta_n} \right) \right] \\ + aL_\Phi \lambda_n^2 \|x_{n+1} - u\|^2 + 2\lambda_n \langle w, u - x_{n+1} \rangle$$

for  $n$  large enough. Since  $\|x_{n+1} - u\|$  converges as  $n \rightarrow \infty$ , it is bounded. Let  $a\|x_{n+1} - u\|^2 \leq M$  for some  $M > 0$  and every  $n$ . Take

$$\varepsilon_n = \frac{d}{2} \lambda_n \beta_n \left[ (\Psi_1 + \Psi_2)^* \left( \frac{4v}{d\beta_n} \right) - \sigma_C \left( \frac{4v}{d\beta_n} \right) \right] + ML_\Phi \lambda_n^2.$$

Assumption  $(H_0)$  iii) and  $(\lambda_n L_\Phi) \in \ell^2$  yield  $\sum_{n \geq 1} \varepsilon_n < +\infty$ . Summing up for  $k = 1, \dots, n$ , we have

$$\|x_{n+1} - u\|^2 - \|x_1 - u\|^2 \leq 2\langle w, \sum_{k=1}^n \lambda_k u \rangle - 2\langle w, \sum_{k=1}^n \lambda_k x_{k+1} \rangle + \sum_{k=1}^n \varepsilon_k.$$

Removing the nonnegative term  $\|x_{n+1} - u\|^2$  and dividing by  $2\tau_n = 2 \sum_{k=1}^n \lambda_k$ , we get

$$(17) \quad \frac{-\|x_1 - u\|^2}{2\tau_n} \leq \langle w, u - \widehat{z}_n \rangle + \frac{1}{2\tau_n} \sum_{k=1}^n \varepsilon_k$$

Passing to the lower limit in (17) and using  $\tau_n \rightarrow \infty$  as  $n \rightarrow \infty$  (because  $\lambda_n \notin \ell^1$ ) we deduce that

$$\liminf_{n \rightarrow \infty} \langle w, u - \widehat{z}_n \rangle \geq 0.$$

If some subsequence  $(\widehat{z}_{n_k})$  converges weakly to  $x_\infty$ , then  $\langle w, u - x_\infty \rangle \geq 0$ . Thus  $x_\infty \in S$ .

For the sequence  $(z_n)$ , we decompose the term  $\langle w, u - x_{n+1} \rangle$  in (16) and write

$$\|x_{n+1} - u\|^2 - \|x_n - u\|^2 \leq \frac{d}{2} \lambda_n \beta_n \left[ (\Psi_1 + \Psi_2)^* \left( \frac{4v}{d\beta_n} \right) - \sigma_C \left( \frac{4v}{d\beta_n} \right) \right] \\ + aL_\Phi \lambda_n^2 \|x_{n+1} - u\|^2 + 2\lambda_n \langle w, x_n - x_{n+1} \rangle + 2\lambda_n \langle w, u - x_n \rangle.$$

Using  $2\lambda_n \langle w, x_n - x_{n+1} \rangle \leq \lambda_n^2 \|w\|^2 + \|x_{n+1} - x_n\|^2$  in the last inequality and proceeding as above we obtain

$$\|x_{n+1} - u\|^2 - \|x_1 - u\|^2 \leq 2\langle w, \sum_{k=1}^n \lambda_k u \rangle - 2\langle w, \sum_{k=1}^n \lambda_k x_k \rangle + \sum_{k=1}^n \zeta_k,$$

where

$$\zeta_n = \frac{d}{2} \lambda_n \beta_n \left[ (\Psi_1 + \Psi_2)^* \left( \frac{4v}{d\beta_n} \right) - \sigma_C \left( \frac{4v}{d\beta_n} \right) \right] + (ML_\Phi + \|w\|^2) \lambda_n^2 + \|x_{n+1} - x_n\|^2.$$

Assumption (H<sub>0</sub>) *iii*), Proposition 17 *ii*) and the additional assumption  $(\lambda_n) \in \ell^2$  give  $\sum_{n \geq 1} \zeta_n < +\infty$  and we conclude as before.  $\blacksquare$

**5.3. Strong convergence: proof of Theorem 2.** In this Section we treat the particular case where additionally the maximal monotone operator  $A$  is strongly monotone, or the potential  $\Phi$  is strongly convex. Recall that  $A$  is strongly monotone with parameter  $\alpha > 0$  if

$$\langle x^* - y^*, x - y \rangle \geq \alpha \|x - y\|^2$$

whenever  $x^* \in Ax$  and  $y^* \in Ay$ . The potential  $\Phi$  is strongly convex if  $\nabla\Phi$  is strongly monotone. Since  $A$ ,  $\nabla\Phi$  and  $N_C$  are monotone, the operator  $\mathbf{T}$  is also strongly monotone whenever  $A$  is strongly monotone or  $\Phi$  is strongly convex. Then, the set  $\mathbf{T}^{-1}0$  reduces to the singleton  $\{u\}$ , for some  $u \in \mathcal{H}$ . By the definition of  $S$ , there exist  $z \in Au$  and  $v \in N_C(u)$  such that  $z + \nabla\Phi(u) + v = 0$ .

The proof of Theorem 2 is a direct consequence of the following reinforced version of Lemma 16:

**Lemma 25.** *There exist  $a, b, c, d, e > 0$  such that, for  $n$  large enough, we have*

$$\begin{aligned} \alpha \lambda_n \|x_{n+1} - u\|^2 + \|x_{n+1} - u\|^2 - \|x_n - u\|^2 + b \|x_{n+1} - x_n\|^2 + c \lambda_n \|\nabla\Phi(x_{n+1}) - \nabla\Phi(u)\|^2 \\ + \frac{d}{2} \lambda_n \beta_n (\Psi_1 + \Psi_2)(x_{n+1}) + e \lambda_n \beta_n \|\nabla\Psi_1(x_n)\|^2 \\ \leq \frac{d}{2} \lambda_n \beta_n \left[ (\Psi_1 + \Psi_2)^* \left( \frac{4v}{d\beta_n} \right) - \sigma_C \left( \frac{4v}{d\beta_n} \right) \right] + a L_\Phi \lambda_n^2 \|x_{n+1} - u\|^2. \end{aligned}$$

**5.4. Weak convergence: proof of Theorem 3.** This section achieves the proof of Theorem 3, that is the weak convergence of the sequence  $(x_n)$  generated by the (SFBP) algorithm, in the special case where  $A = \partial\Phi_2$  is the subdifferential of a proper lower-semicontinuous convex function  $\Phi_2 : \mathcal{H} \rightarrow \mathbf{R} \cup \{+\infty\}$ . Writing  $\Phi_1$  instead of  $\Phi$ , for the sake of symmetry, the (SFBP) algorithm takes the form

$$(18) \quad \begin{cases} x_1 & \in \mathcal{H}, \\ x_{n+1} & = (I + \lambda_n \partial\Phi_2 + \lambda_n \beta_n \partial\Psi_2)^{-1}(x_n - \lambda_n \nabla\Phi_1(x_n) - \lambda_n \beta_n \nabla\Psi_1(x_n)) \quad \forall n \geq 1. \end{cases}$$

Since  $\partial\Phi_2 + \nabla\Phi_1 + N_C$  is maximal monotone, the solution set  $S$  is equal to

$$S = (\partial\Phi_2 + \nabla\Phi_1 + N_C)^{-1}(0) = \text{Argmin}\{\Phi_1(u) + \Phi_2(u) : u \in \text{Argmin}\Psi_1 \cap \text{Argmin}\Psi_2\}.$$

We prove the weak convergence of the sequence  $(x_n)$  generated by algorithm (18) to some point in  $S$  using Opial-Passty's Lemma 15. The first assumption is satisfied from Proposition 17 *i*). The second assumption, that every weak cluster point of  $(x_n)$  belongs to  $S$ , will be verified in three different cases *(i)*, respectively *(ii)* and *(iii)*, in Subsection 5.4.1, respectively Subsection 5.4.2. Finally, in Subsection 5.4.3 we finish the proof of Theorem 3 with the minimizing property of the sequence.

5.4.1. *Weak convergence in case (i): bounded inf-compactness.* Denote by  $\text{dist}(\cdot, S)$  the distance function to the closed convex set  $S$  and set  $d(x) = \frac{1}{2}\text{dist}(x, S)^2$ . The function  $d$  is convex and differentiable,  $\nabla d(x) = x - P_S(x)$  where  $P_S$  denotes the projection onto  $S$ .

The proof goes along the same lines as that of [5, Theorem 16]. In the next lemma, we prove that  $\lim_{n \rightarrow \infty} d(x_n) = 0$ . By the weak lower semicontinuity of the convex function  $d$ , it implies that every weak cluster point of  $(x_n)$  lies in  $S$ . Thus  $(x_n)$  satisfies the second assumption of Opial-Passty's Lemma, and we deduce the weak convergence to some point in  $S$ . Besides, since  $\liminf_{n \rightarrow \infty} \lambda_n \beta_n > 0$  by assumption, Proposition 17 iii) yields  $\lim_{n \rightarrow \infty} (\Psi_1 + \Psi_2)(x_n) = 0$ . If we additionally assume that  $(\Psi_1 + \Psi_2)$  is boundedly inf-compact, the bounded sequence  $(x_n)$  is also relatively compact. Hence its weak convergence implies its strong convergence. This achieves the proof of the first part of Theorem 3 in Case (i).

**Lemma 26.** *Under the assumptions of Theorem 3 (i), let  $(x_n)$  be a sequence generated by the (SFBP) algorithm. Then*

$$\lim_{n \rightarrow \infty} d(x_n) = 0.$$

**Proof.** We reformulate (18) as

$$(19) \quad x_n - x_{n+1} = \lambda_n \nabla \Phi_1(x_n) + \lambda_n \beta_n \nabla \Psi_1(x_n) + \lambda_n v_{n+1} + \lambda_n \beta_n w_{n+1},$$

where  $v_{n+1} \in \partial \Phi_2(x_{n+1})$  and  $w_{n+1} \in \partial \Psi_2(x_{n+1})$ . The convexity of  $d$  with (19) yields

$$(20) \quad \begin{aligned} d(x_n) &\geq d(x_{n+1}) + \langle x_{n+1} - P_S(x_{n+1}), x_n - x_{n+1} \rangle \\ &= d(x_{n+1}) + \lambda_n \langle \nabla \Phi_1(x_n) + v_{n+1}, x_{n+1} - P_S(x_{n+1}) \rangle \\ &\quad + \lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_{n+1} - P_S(x_{n+1}) \rangle + \lambda_n \beta_n \langle w_{n+1}, x_{n+1} - P_S(x_{n+1}) \rangle. \end{aligned}$$

We treat each term on the right-hand side of (20). Let

$$\alpha = \min\{(\Phi_1 + \Phi_2)(z); z \in C\}.$$

Firstly for  $\lambda_n \langle \nabla \Phi_1(x_n) + v_{n+1}, x_{n+1} - P_S(x_{n+1}) \rangle$ . Since  $\Phi_1$  is convex we have

$$(21) \quad \begin{aligned} \Phi_1(P_S x_{n+1}) &\geq \Phi_1(x_n) + \langle \nabla \Phi_1(x_n), P_S x_{n+1} - x_n \rangle \\ &= \Phi_1(x_n) + \langle \nabla \Phi_1(x_n), P_S x_{n+1} - x_{n+1} \rangle + \langle \nabla \Phi_1(x_n), x_{n+1} - x_n \rangle. \end{aligned}$$

From Descent Lemma (Lemma 21) we have

$$\Phi_1(x_{n+1}) \leq \Phi_1(x_n) + \langle \nabla \Phi_1(x_n), x_{n+1} - x_n \rangle + \frac{L_{\Phi_1}}{2} \|x_{n+1} - x_n\|^2.$$

Using this in (21) it follows that

$$(22) \quad \Phi_1(P_S x_{n+1}) \geq \Phi_1(x_{n+1}) - \frac{L_{\Phi_1}}{2} \|x_{n+1} - x_n\|^2 + \langle \nabla \Phi_1(x_n), P_S x_{n+1} - x_{n+1} \rangle.$$

On the other hand, the subdifferential inequality for  $\Phi_2$  writes

$$(23) \quad \Phi_2(P_S x_{n+1}) \geq \Phi_2(x_{n+1}) + \langle v_{n+1}, P_S x_{n+1} - x_{n+1} \rangle.$$

Noting that  $\Phi_1(P_S x_{n+1}) + \Phi_2(P_S x_{n+1}) = \alpha$  and adding (22) and (23) we get

$$(24) \quad \langle \nabla \Phi_1(x_n) + v_{n+1}, P_S x_{n+1} - x_{n+1} \rangle \leq \alpha - (\Phi_1 + \Phi_2)(x_{n+1}) + \frac{L_{\Phi_1}}{2} \|x_{n+1} - x_n\|^2.$$

Secondly for  $\lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_{n+1} - P_S(x_{n+1}) \rangle$ . If  $\Psi_1 = 0$  then it is equal to zero. If  $\Psi_1 \neq 0$ , since  $\nabla \Psi_1(P_S x_{n+1}) = 0$ , the cocoercivity of  $\nabla \Psi_1$  implies

$$\langle \nabla \Psi_1(x_n), P_S x_{n+1} - x_n \rangle \leq -\frac{1}{L_{\Psi_1}} \|\nabla \Psi_1(x_n)\|^2.$$

Adding the last inequality and

$$\langle \nabla \Psi_1(x_n), x_n - x_{n+1} \rangle \leq \frac{1}{L_{\Psi_1}} \|\nabla \Psi_1(x_n)\|^2 + \frac{L_{\Psi_1}}{4} \|x_{n+1} - x_n\|^2$$

we deduce

$$(25) \quad \langle \nabla \Psi_1(x_n), P_S x_{n+1} - x_{n+1} \rangle \leq \frac{L_{\Psi_1}}{4} \|x_{n+1} - x_n\|^2.$$

Thirdly for  $\lambda_n \beta_n \langle w_{n+1}, x_{n+1} - P_S(x_{n+1}) \rangle$ . Since  $w_{n+1} \in \partial \Psi_2(w_{n+1})$  and  $0 \in \partial \Psi_2(P_S x_{n+1})$ , the monotonicity of  $\partial \Psi_2$  implies

$$(26) \quad \langle w_{n+1}, x_{n+1} - P_S x_{n+1} \rangle \geq 0.$$

Combining (24), (25) and (26) in (20), and since  $\limsup_{n \rightarrow \infty} (L_{\Psi_1} \lambda_n \beta_n) < 2$ , we deduce that

$$\begin{aligned} d(x_{n+1}) - d(x_n) + \lambda_n [(\Phi_1 + \Phi_2)(x_{n+1}) - \alpha] &\leq \lambda_n \frac{L_{\Phi_1}}{2} \|x_{n+1} - x_n\|^2 + \lambda_n \beta_n \frac{L_{\Psi_1}}{4} \|x_{n+1} - x_n\|^2 \\ &\leq \frac{1}{2} \|x_{n+1} - x_n\|^2, \end{aligned}$$

for  $n$  large enough. The remainder of the proof is an adaptation of the proof of [5, Theorem 16]: considering that  $x_{n+1}$  may not lie in  $C$ , and we may not have  $(\Phi_1 + \Phi_2)(x_{n+1}) - \alpha \geq 0$  for every  $n \in \mathbf{N}$ , it is achieved by studying separately the two cases:

Case I: There exists  $n_0 \in \mathbf{N}$  such that  $(\Phi_1 + \Phi_2)(x_n) \geq \alpha$  for all  $n \geq n_0$ .

Case II: For each  $n \in \mathbf{N}$  there exists  $n' > n$  such that  $(\Phi_1 + \Phi_2)(x_{n'}) < \alpha$ . ■

**5.4.2. Weak convergence in cases (ii) and (iii): bounded increase of the sequence  $(\beta_n)$  and the unconstrained case.** As before, it suffices to prove that every weak cluster point of the sequence  $(x_n)$  generated by the (SFBP) algorithm lies in  $S$  to deduce its weak convergence to a point in  $S$ . We decompose the proof in several lemmas. Let us introduce the penalized functions  $\Omega_n$  and  $\tilde{\Omega}_n$  defined on  $\mathcal{H}$  by

$$\Omega_n = \Phi_1 + \beta_n \Psi_1 \quad \text{and} \quad \tilde{\Omega}_n = \Phi_2 + \beta_n \Psi_2.$$

Being the sum of two smooth functions whose gradient is lipschitz-continuous,  $\Omega_n$  is in his turn a smooth function whose gradient is lipschitz-continuous with constant  $L_n = L_{\Phi_1} + \beta_n L_{\Psi_1}$ . If  $\Psi_1 = 0$ ,  $\Omega_n$  reduces to  $\Phi_1$  and  $L_n = L_{\Phi_1}$ .

**Lemma 27.** *Assume that  $(H_0)$  hold with  $A = \partial \Phi_2$ ,  $(\lambda_n L_{\Phi}) \in \ell^2$  and let  $(x_n)$  be a sequence generated by the (SFBP) algorithm. Then the following holds:*

i) *For every  $n \geq 1$  the penalized functions verify:*

$$\begin{aligned} \left[ \Omega_{n+1}(x_{n+1}) + \tilde{\Omega}_{n+1}(x_{n+1}) \right] - \left[ \Omega_n(x_n) + \tilde{\Omega}_n(x_n) \right] \\ + \left[ \frac{1}{\lambda_n} - \frac{L_n}{2} \right] \|x_{n+1} - x_n\|^2 \leq (\beta_{n+1} - \beta_n)(\Psi_1 + \Psi_2)(x_{n+1}). \end{aligned}$$

ii) *If  $\beta_{n+1} - \beta_n \leq K \lambda_{n+1} \beta_{n+1}$ , or if  $\Psi_1 = \Psi_2 = 0$ , then the sequence  $(\Omega_n(x_n) + \tilde{\Omega}_n(x_n))$  converges as  $n \rightarrow \infty$ .*

**Proof of Part i).** Apply Descent Lemma 21 to obtain

$$\Omega_n(x_{n+1}) - \Omega_n(x_n) \leq \langle \nabla \Omega_n(x_n), x_{n+1} - x_n \rangle + \frac{L_n}{2} \|x_{n+1} - x_n\|^2.$$



Using  $\Omega_n(x_{n+1}) = \Omega_{n+1}(x_{n+1}) - (\beta_{n+1} - \beta_n)\Psi_1(x_{n+1})$ , it follows that

$$(27) \quad \Omega_{n+1}(x_{n+1}) - \Omega_n(x_n) \leq \langle \nabla \Omega_n(x_n), x_{n+1} - x_n \rangle + \frac{L_n}{2} \|x_{n+1} - x_n\|^2 + (\beta_{n+1} - \beta_n)\Psi_1(x_{n+1}).$$

On the other hand, remark that

$$(28) \quad \tilde{\Omega}_{n+1}(x_{n+1}) - \tilde{\Omega}_n(x_n) = \Phi_2(x_{n+1}) - \Phi_2(x_n) + \beta_n [\Psi_2(x_{n+1}) - \Psi_2(x_n)] + (\beta_{n+1} - \beta_n)\Psi_2(x_{n+1}).$$

Recall the following formulation of the (SFBP) algorithm

$$x_n - x_{n+1} = \lambda_n \nabla \Phi_1(x_n) + \lambda_n \beta_n \nabla \Psi_1(x_n) + \lambda_n v_{n+1} + \lambda_n \beta_n w_{n+1},$$

with  $v_{n+1} \in \partial \Phi_2(x_{n+1})$  and  $w_{n+1} \in \partial \Psi_2(x_{n+1})$ . The subdifferential inequality at  $x_{n+1}$  of  $\Phi_2$  and  $\Psi_2$ , respectively gives

$$\Phi_2(x_{n+1}) - \Phi_2(x_n) \leq \langle v_{n+1}, x_{n+1} - x_n \rangle \quad \text{and} \quad \Psi_2(x_{n+1}) - \Psi_2(x_n) \leq \langle w_{n+1}, x_{n+1} - x_n \rangle.$$

Replacing this in (28) we obtain

$$\tilde{\Omega}_{n+1}(x_{n+1}) - \tilde{\Omega}_n(x_n) \leq \langle v_{n+1} + \beta_n w_{n+1}, x_{n+1} - x_n \rangle + (\beta_{n+1} - \beta_n)\Psi_2(x_{n+1}).$$

Adding (27) and the last inequality we deduce that

$$(29) \quad \begin{aligned} & \left[ \Omega_{n+1}(x_{n+1}) + \tilde{\Omega}_{n+1}(x_{n+1}) \right] - \left[ \Omega_n(x_n) + \tilde{\Omega}_n(x_n) \right] \leq \langle \nabla \Omega_n(x_n) + v_{n+1} + \beta_n w_{n+1}, x_{n+1} - x_n \rangle \\ & \quad + \frac{L_n}{2} \|x_{n+1} - x_n\|^2 + (\beta_{n+1} - \beta_n)(\Psi_1 + \Psi_2)(x_{n+1}). \end{aligned}$$

Therefore, just substitute the equality  $\nabla \Omega_n(x_n) + v_{n+1} + \beta_n w_{n+1} = -\frac{x_{n+1} - x_n}{\lambda_n}$  to conclude *i*).

**Proof of Part ii).** Since  $L_{\Psi_1} \lambda_n \beta_n < 2$  for  $n$  large enough, from *i*) we have

$$(30) \quad \begin{aligned} & \left[ \Omega_{n+1}(x_{n+1}) + \tilde{\Omega}_{n+1}(x_{n+1}) \right] - \left[ \Omega_n(x_n) + \tilde{\Omega}_n(x_n) \right] \\ & \leq (\beta_{n+1} - \beta_n)(\Psi_1 + \Psi_2)(x_{n+1}) + \frac{L_{\Phi_1}}{2} \|x_{n+1} - x_n\|^2. \end{aligned}$$

Take an element  $u \in S$  and  $z \in \partial \Phi_2(u)$ . Write the subdifferential inequality at  $u$  for  $\Phi_1$  and  $\Phi_2$  to obtain

$$(31) \quad \begin{aligned} \Omega_n(x_n) + \tilde{\Omega}_n(x_n) & \geq \Phi_1(x_n) + \Phi_2(x_n) \\ & \geq \Phi_1(u) + \Phi_2(u) + \langle \nabla \Phi_1(u) + z, x_n - u \rangle. \end{aligned}$$

Since  $(x_n)$  is bounded by Proposition 17 i), the sequence  $(\Omega_n(x_n) + \tilde{\Omega}_n(x_n))$  is bounded from below. Now, if  $\beta_{n+1} - \beta_n \leq K \lambda_{n+1} \beta_{n+1}$ , or if  $\Psi_1 = \Psi_2 = 0$ , the right-hand side of (30) is summable from Proposition 17 ii). This implies the convergence of the sequence  $(\Omega_n(x_n) + \tilde{\Omega}_n(x_n))$ . ■

**Lemma 28.** Assume  $(H_0)$ ,  $(\lambda_n L_{\Phi}) \in \ell^2$ , and additionnally  $\lambda_n$  is bounded and  $\beta_{n+1} - \beta_n \leq K \lambda_{n+1} \beta_{n+1}$ , respectively  $\Psi_1 = \Psi_2 = 0$ . Let  $(x_n)$  be a sequence generated by the (SFBP) algorithm. Then, for every  $u \in S$  we have

$$(32) \quad \sum_{n \geq 1} \lambda_n \left[ \Omega_{n+1}(x_{n+1}) + \tilde{\Omega}_{n+1}(x_{n+1}) - \Phi_1(u) - \Phi_2(u) \right] < +\infty \quad (\text{possibly } -\infty).$$

**Proof.** We write

$$(33) \quad 2\lambda_n \left[ \Omega_{n+1}(x_{n+1}) + \tilde{\Omega}_{n+1}(x_{n+1}) - \Phi_1(u) - \Phi_2(u) \right] = \\ 2\lambda_n [\Phi_1(x_{n+1}) - \Phi_1(u)] + 2\lambda_n [\Phi_2(x_{n+1}) - \Phi_2(u)] + 2\lambda_n \beta_{n+1} (\Psi_1 + \Psi_2)(x_{n+1}).$$

The (SFBP) algorithm writes

$$v_{n+1} = \frac{x_n - x_{n+1}}{\lambda_n} - \nabla \Phi_1(x_n) - \beta_n \nabla \Psi_1(x_n) - \beta_n w_{n+1}$$

with  $v_{n+1} \in \partial \Phi_2(x_{n+1})$  and  $w_{n+1} \in \partial \Psi_2(x_{n+1})$ . The subdifferential inequality of  $\Phi_1$  and  $\Phi_2$ , respectively gives

$$\Phi_1(x_{n+1}) - \Phi_1(u) \leq \langle \nabla \Phi_1(x_{n+1}), x_{n+1} - u \rangle \quad \text{and} \quad \Phi_2(x_{n+1}) - \Phi_2(u) \leq \langle v_{n+1}, x_{n+1} - u \rangle.$$

Thus

$$2\lambda_n \left[ \Omega_{n+1}(x_{n+1}) + \tilde{\Omega}_{n+1}(x_{n+1}) - \Phi_1(u) - \Phi_2(u) \right] \leq \\ 2\lambda_n \langle \nabla \Phi_1(x_{n+1}), x_{n+1} - u \rangle + 2\lambda_n \langle v_{n+1}, x_{n+1} - u \rangle + 2\lambda_n \beta_{n+1} (\Psi_1 + \Psi_2)(x_{n+1}).$$

Write  $2\langle x_{n+1} - x_n, x_{n+1} - u \rangle = \|x_{n+1} - x_n\|^2 + \|x_{n+1} - u\|^2 - \|x_n - u\|^2$  and deduce

$$(34) \quad 2\lambda_n \left[ \Omega_{n+1}(x_{n+1}) + \tilde{\Omega}_{n+1}(x_{n+1}) - \Phi_1(u) - \Phi_2(u) \right] \leq -\|x_{n+1} - x_n\|^2 - \|x_{n+1} - u\|^2 + \|x_n - u\|^2 \\ + 2\lambda_n \langle \nabla \Phi_1(x_{n+1}) - \nabla \Phi_1(x_n), x_{n+1} - u \rangle + 2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), u - x_{n+1} \rangle \\ + 2\lambda_n \beta_n \langle w_{n+1}, u - x_{n+1} \rangle + 2\lambda_n \beta_{n+1} \Psi_2(x_{n+1}) + 2\lambda_n \beta_{n+1} \Psi_1(x_{n+1}).$$

We now treat each term on the right-hand side of (34).

For the term.  $2\lambda_n \langle \nabla \Phi_1(x_{n+1}) - \nabla \Phi_1(x_n), x_{n+1} - u \rangle$  on the right-hand side of (34), use the Cauchy-Schwartz inequality to write

$$2\lambda_n \langle \nabla \Phi_1(x_{n+1}) - \nabla \Phi_1(x_n), x_{n+1} - u \rangle \leq 2\lambda_n \|\nabla \Phi_1(x_{n+1}) - \nabla \Phi_1(x_n)\| \|x_{n+1} - u\| \\ \leq 2\lambda_n L_{\Phi_1} \|x_{n+1} - x_n\| \|x_{n+1} - u\| \\ \leq L_{\Phi_1} \|x_{n+1} - x_n\|^2 + \lambda_n^2 L_{\Phi_1} \|x_{n+1} - u\|^2$$

Since  $\sum_{n \geq 1} \|x_{n+1} - x_n\|^2 < +\infty$ ,  $\sum_{n \geq 1} \lambda_n^2 L_{\Phi_1} < +\infty$  and  $(\|x_n - u\|)$  is bounded, we deduce that

$$(35) \quad \sum_{n \geq 1} 2\lambda_n \langle \nabla \Phi_1(x_{n+1}) - \nabla \Phi_1(x_n), x_{n+1} - u \rangle < +\infty.$$

For the term.  $2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), u - x_{n+1} \rangle$ , write

$$(36) \quad 2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), u - x_{n+1} \rangle = 2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), u - x_n \rangle + 2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_n - x_{n+1} \rangle.$$

On one hand, the monotonicity of the gradient and the fact  $u \in C$  imply

$$(37) \quad \langle \nabla \Psi_1(x_n), u - x_n \rangle \leq 0.$$

On the other hand we have

$$2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_n - x_{n+1} \rangle \leq \lambda_n^2 \beta_n^2 \|\nabla \Psi_1(x_n)\|^2 + \|x_{n+1} - x_n\|^2.$$

Therefore Proposition 17 ii) and the bound  $\lambda_n \beta_n < \frac{2}{L_{\Psi_1}}$ , if  $L_{\Psi_1} \neq 0$  and for  $n$  large enough, yield

$$(38) \quad \sum_{n \geq 1} \lambda_n \beta_n \langle \nabla \Psi_1(x_n), x_n - x_{n+1} \rangle < +\infty.$$

Combining (37) and (38) in (36) we conclude

$$(39) \quad \sum_{n \geq 1} \lambda_n \beta_n \langle \nabla \Psi_1(x_n), u - x_{n+1} \rangle < +\infty.$$

If  $L_{\Psi_1} = 0$ , use the Cauchy-Schwartz inequality and write

$$2\lambda_n \beta_n \langle \nabla \Psi_1(x_n), u - x_{n+1} \rangle \leq 2\lambda_n \beta_n \|\nabla \Psi_1(x_n)\| \|u - x_{n+1}\|,$$

which is summable, since  $\|u - x_{n+1}\|$  is bounded,  $\|\nabla \Psi_1(x_n)\| = a \|\nabla \Psi_1(x_n)\|^2$  with  $a > 0$ , and in view of Proposition 17 ii).

For the term  $2\lambda_n \beta_n \langle w_{n+1}, u - x_{n+1} \rangle + 2\lambda_n \beta_{n+1} \Psi_2(x_{n+1})$ , if  $\Psi_2 = 0$ , its value is zero. Otherwise, we assume  $\beta_{n+1} - \beta_n \leq K\lambda_{n+1}\beta_{n+1}$  and use the subdifferential inequality for  $\Psi_2$  at  $x_{n+1}$  and  $u$

$$\langle w_{n+1}, u - x_{n+1} \rangle \leq -\Psi_2(x_{n+1})$$

to write

$$2\lambda_n \beta_n \langle w_{n+1}, u - x_{n+1} \rangle + 2\lambda_n \beta_{n+1} \Psi_2(x_{n+1}) \leq 2K\lambda_n \lambda_{n+1} \beta_{n+1} \Psi_2(x_{n+1}).$$

Since  $\lambda_n$  is bounded by assumption and  $\sum_{n \geq 1} \lambda_{n+1} \beta_{n+1} \Psi_2(x_{n+1}) < \infty$  by Proposition 17 iii), it follows that

$$(40) \quad \sum_{n \geq 1} [\lambda_n \beta_n \langle w_{n+1}, u - x_{n+1} \rangle + \lambda_n \beta_{n+1} \Psi_2(x_{n+1})] < +\infty.$$

For the remaining term  $2\lambda_n \beta_{n+1} \Psi_1(x_{n+1})$ , it is equal to zero if  $\Psi_1 = 0$ . Otherwise, we assume  $\beta_{n+1} - \beta_n \leq K\lambda_{n+1}\beta_{n+1}$ , and we write

$$\lambda_n \beta_{n+1} \Psi_1(x_{n+1}) = \lambda_n \beta_n \Psi_1(x_{n+1}) + \lambda_n (\beta_{n+1} - \beta_n) \Psi_1(x_{n+1}).$$

Noting that,  $\lambda_n \beta_n < 2/L_{\Psi_1}$  for  $n$  large enough,  $\lambda_n$  is bounded by assumption and that  $\lambda_n (\beta_{n+1} - \beta_n) \leq K\lambda_n \lambda_{n+1} \beta_{n+1}$ , Proposition 17 iii) yields  $\sum_{n \geq 1} \lambda_n \beta_n \Psi_1(x_{n+1}) < +\infty$  and  $\sum_{n \geq 1} \lambda_n (\beta_{n+1} - \beta_n) \Psi_1(x_{n+1}) < +\infty$ . We then deduce that

$$(41) \quad \sum_{n \geq 1} \lambda_n \beta_{n+1} \Psi_1(x_{n+1}) < +\infty.$$

Finally we conclude (32) from (34) by using (35), (39), (40), (41) and the fact that

$$(42) \quad \sum_{n \geq 1} \|x_n - u\|^2 - \|x_{n+1} - u\|^2 \leq \|x_1 - u\|^2.$$

■

**Lemma 29.** Assume  $(H_0)$  and  $(\lambda_n L_\Phi) \in \ell^2$ . Assume moreover that one of the following conditions holds:

- (i)  $\liminf_{n \rightarrow \infty} \lambda_n \beta_n > 0$ ,  $(\lambda_n)$  is bounded and  $\beta_{n+1} - \beta_n \leq K\lambda_{n+1}\beta_{n+1}$  for some  $K > 0$ .
- (ii)  $\Psi_1 = \Psi_2 = 0$ .

Let  $(x_{n_k})$  be a subsequence of  $(x_n)$  that converges weakly to some  $x_\infty$  as  $n \rightarrow \infty$ . Then

$$x_\infty \in S = \text{Argmin}\{\Phi_1(x) + \Phi_2(x) : x \in \text{Argmin}(\Psi_1 + \Psi_2)\}.$$

**Proof.** Since  $\sum_{n \geq 1} \lambda_n = +\infty$  by the last statement of Hypotheses  $(H_0)$ , Lemmas 27 ii) and 28 together imply

$$(43) \quad \lim_{n \rightarrow \infty} [\Omega_n(x_n) + \tilde{\Omega}_n(x_n)] \leq \Phi_1(u) + \Phi_2(u) \quad \forall u \in S.$$

Now, in view of (43), the weak lower-semicontinuity of  $\Phi_1$  and  $\Phi_2$  yields

$$\begin{aligned}
 (44) \quad \Phi_1(x_\infty) + \Phi_2(x_\infty) &\leq \liminf_{k \rightarrow \infty} \Phi_1(x_{n_k}) + \liminf_{k \rightarrow \infty} \Phi_2(x_{n_k}) \\
 &\leq \liminf_{k \rightarrow \infty} \Omega_{n_k}(x_{n_k}) + \liminf_{k \rightarrow \infty} \tilde{\Omega}_{n_k}(x_{n_k}) \\
 &\leq \liminf_{k \rightarrow \infty} \left[ \Omega_{n_k}(x_{n_k}) + \tilde{\Omega}_{n_k}(x_{n_k}) \right] \\
 &= \lim_{n \rightarrow \infty} \left[ \Omega_n(x_n) + \tilde{\Omega}_n(x_n) \right] \\
 &\leq \Phi_1(u) + \Phi_2(u) \quad \forall u \in S.
 \end{aligned}$$

Under Assumption (i),  $x_\infty$  belongs to  $C$  by Proposition 17 *iii*). Under Assumption (ii),  $C = \mathcal{H}$ . Thus  $x_\infty \in S$  and every weak cluster point of  $(x_n)$  lies in  $S$ .  $\blacksquare$

5.4.3. *Minimizing property in cases (ii) and (iii).* The first part of Theorem 3, the weak convergence of the sequence  $(x_n)$ , is proved in Subsection 5.4.1 in Case (i), and in Subsection 5.4.2 in Cases (ii) and (iii). For the second part, recalling that  $\nabla \Phi_1(u) = -p$  and  $z \in \partial \Phi_2(u)$ , the subdifferential inequality of  $\Phi_1 + \Phi_2$  at  $u \in S$  yields

$$(45) \quad (\Phi_1 + \Phi_2)(x_n) \geq (\Phi_1 + \Phi_2)(u) + \langle z - p, x_n - u \rangle.$$

Passing to the lower limit in (45), using  $p - z \in N_C(u)$  and the fact that  $(x_n)$  weakly converges to a point in  $S$  (First part of Theorem 3), it follows that

$$\liminf_{n \rightarrow \infty} (\Phi_1 + \Phi_2)(x_n) \geq (\Phi_1 + \Phi_2)(u).$$

On the other hand, using (43) we have

$$\limsup_{n \rightarrow \infty} (\Phi_1 + \Phi_2)(x_n) \leq \lim_{n \rightarrow \infty} \left[ \Omega_n(x_n) + \tilde{\Omega}_n(x_n) \right] \leq \Phi_1(u) + \Phi_2(u).$$

Combining the last two inequalities with the fact that  $u \in S$ , the second part of Theorem 3 directly follows.  $\blacksquare$

## REFERENCES

- [1] F. Álvarez and J. Peypouquet, *Asymptotic almost-equivalence of Lipschitz evolution systems in Banach spaces*, Nonlinear Anal. 73 (2010), no. 9, 3018–3033.
- [2] H. Attouch and R. Cominetti, *A dynamical approach to convex minimization coupling approximation with the steepest descent method*, J. Differential Equations, 128 (1996), 519–540.
- [3] H. Attouch and M.-O. Czarnecki, *Asymptotic behavior of coupled dynamical systems with multiscale aspects*, J. Differ. Equations 248 (2010) no. 6, 1315–1344.
- [4] H. Attouch, M.-O. Czarnecki and J. Peypouquet, *Prox-penalization and splitting methods for constrained variational problems*, SIAM J. Optim., 21 (2011) no. 1, 149–173.
- [5] H. Attouch, M.-O. Czarnecki and J. Peypouquet, *Coupling forward-backward with penalty schemes and parallel splitting for constrained variational inequalities*, SIAM J. Optim. 21 (2011) no. 4, 1251–1274.
- [6] J.-B. Baillon, *Un exemple concernant le comportement asymptotique de la solution du problème  $du/dt + \partial \phi(u) = 0$* , J. Functional Anal. 28, (1978), 369–376.
- [7] J.-B. Baillon, *Comportement asymptotique des contractions et semi-groupes de contractions - equations de schroedinger non lineaires et divers*, Thèse, Paris VI, 1978.
- [8] J.-B. Baillon and H. Brézis, *Une remarque sur le comportement asymptotique des semi-groupes non linéaires*, Houston J. Math. 2, (1976), 5–7.
- [9] J.-B. Baillon and R. Cominetti, *A convergence result for nonautonomous subgradient evolution equations and its application to the steepest descent exponential penalty trajectory in linear programming*, J. Funct. Anal. 187 (2001), 263–273.

- [10] J.-B. Baillon and G. Haddad, *Quelques propriétés des opérateurs angle-bornés et  $n$ -cycliquement monotones*, Israel J. Math. 26 (1977), no. 2, 137–150.
- [11] D. Bertsekas, *Nonlinear programming*. Athena Scientific, Belmont MA, 2nd Printing (2003).
- [12] R. I. Boş, E. R. Csetnek *Forward-Backward and Tseng's type penalty schemes for monotone inclusion problems*, Set-Valued and Variational Analysis 22(2) (2014), 313–331
- [13] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North Holland Publishing Company, Amsterdam, 1973.
- [14] H. Brézis and P.-L. Lions, *Produits infinis de résolvantes*, Israel J. Math. 29 (1978), 329–345.
- [15] R.-E. Bruck, *Asymptotic convergence of nonlinear contraction semigroups in Hilbert spaces*, J. Funct. Anal. 18 (1975), 15–26.
- [16] R.-E. Bruck, *On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space*, J. Math. Anal. Appl. 61 (1977), 159–164.
- [17] R. Burachik, L.-M. Grana Drummond, A.-N. Iseum and B.-F. Svaiter, *Full convergence of the steepest Descent method with inexact line searches*, Optimization 32 (1995), 137–146.
- [18] A. Cabot, *The steepest descent dynamical system with control. Applications to constrained minimization*, ESAIM Control Optim. Calc. Var. 10 (2004), 243–258.
- [19] A. Cabot, *Proximal point algorithm controlled by a slowly vanishing term. Applications to hierarchical minimization*, SIAM J. Optim. 15 (2005), no. 8, 1207–1223.
- [20] A.-L. Cauchy, *Méthode générale pour la résolution des systèmes d'équations simultanées*, C. R. Acad. Sci. Paris 25 (1847), 536–538.
- [21] P.-L. Combettes, *Solving monotone inclusions via compositions of nonexpansive averaged operators*, Optimization 53 (2004), no. 5-6, 475–504.
- [22] L.-C. Evans, *Partial differential equations*, Second edition, Graduate Studies in Mathematics 19, American Mathematical Society, Providence, RI, 2010.
- [23] O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim. 29 (1991), no. 2, 403–419.
- [24] P.-L. Lions, *Une methode iterative de resolution d'une inequation variationnelle*, Israel J. Math. 31 (1978), no. 2, 204–208.
- [25] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, RAIRO 4 (1970), 154–159.
- [26] B. Martinet, *Détermination approchée d'un point fixe d'une application pseudo-contractante*, C.R. Acad. Sci. Paris. 274 (1972), 163–165.
- [27] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France 93 (1965), 273–299.
- [28] N. Noun, *Convergence et stabilisation de systèmes dynamiques couplés et multi-échelles vers des équilibres sous contraintes ; application à l'optimisation hiérarchique*, PhD Thesis, Université Montpellier 2 and Université Libanaise à Beyrouth, 2013.
- [29] N. Noun and J. Peypouquet, *Forward-Backward-Penalty scheme for constrained convex minimization without inf-compactness*, J. Optim. Theory Appl. 158 (2013), no. 3, 787–795.
- [30] Z. Opial, *Weak Convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc. 73 (1967), 591–597.
- [31] G. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, J. Math. Anal. Appl. 72 (1979), no. 2, 383–390.
- [32] J. Peypouquet, *Optimización y sistemas dinámicos*, Ediciones IVIC, Caracas, 2013.
- [33] J. Peypouquet, *Coupling the gradient method with a general exterior penalization scheme for convex minimization*, J. Optim. Theory Appl. 153 (2012), no. 1, 123–138.
- [34] J. Peypouquet and S. Sorin, *Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time*, J. Convex Anal. 17 (2010), 1113–1163.
- [35] R.-T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim. 14 (1976), no. 5, 877–897.

INSTITUT DE MATHÉMATIQUES ET MODÉLISATION DE MONTPELLIER, UMR 5149 CNRS, UNIVERSITÉ MONTPELLIER 2, PLACE EUGÈNE BATAILLON, 34095 MONTPELLIER CEDEX 5, FRANCE  
*E-mail address:* marco@univ-montp2.fr, nahla.noun@yahoo.fr

DEPARTAMENTO DE MATEMÁTICA, UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA, AVENIDA ESPAÑA 1680, VALPARAÍSO, CHILE.  
*E-mail address:* juan.peypouquet@usm.cl